Bankhofer, Udo:

Unvollständige Daten- und Distanzmatrizen in der multivariaten Datenanalyse

DOI: 10.22032/dbt.40346

URN: urn:nbn:de:gbv:ilm1-2020300018

Retrodigitalisierung der gleichnamigen Ausgabe:

Erschienen als:	Bankhofer, Udo: Unvollständige Daten- und Distanzmatrizen in der multivariaten Datenanalyse / Udo Bankhofer. - Bergisch Gladbach [u.a.] : Eul, 1995. - VI, 220 S. ISBN 3-89012-458-5 (Reihe: Quantitative Ökonomie ; 64) Zugl.: Augsburg, Univ., Diss., 1995
Digitalisierung durch:	Universitätsbibliothek Ilmenau / ilmedia
Digitalisierungsjahr:	2020
Scanformat:	TIFF, 600 DPI, 8 BPP

Reihe: Quantitative Ökonomie · Band 64
Herausgegeben von Eckart Bomsdorf, Wim Kösters und Winfried Matthes

Udo Bankhofer

Unvollständige Daten- und Distanzmatrizen in der Multivariaten Datenanalyse

Verlag Josef Eul
Bergisch Gladbach · Köln

Reihe: Quantitative Ökonomie · Band 64

Herausgegeben von Prof. Dr. Eckart Bomsdorf, Köln; Prof. Dr. Wim Kösters, Bochum und Prof. Dr. Winfried Matthes, Wuppertal

Dr. Udo Bankhofer

Unvollständige Daten- und Distanzmatrizen in der Multivariaten Datenanalyse

Verlag Josef Eul
Bergisch Gladbach · Köln

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Bankhofer, Udo:
Unvollständige Daten- und Distanzmatrizen in der Multivariaten Datenanalyse / Udo Bankhofer. – Bergisch Gladbach ; Köln : Eul, 1995.
(Reihe: Quantitative Ökonomie ; Bd. 64)
Zugl.: Augsburg, Univ., Diss., 1995
ISBN 3-89012-458-5
NE: GT

Josef Eul Verlag GmbH
Postfach 10 06 56
51406 Bergisch Gladbach

Printed in Germany
Druck: Rosch-Buch, Hallstadt

Gedruckt auf säurefreiem und 100% chlorfrei gebleichtem Papier

Vorwort

Die vorliegende Arbeit beschäftigt sich mit dem Problem unvollständiger Daten- und Distanzmatrizen in der Multivariaten Datenanalyse. Im Fall fehlender Werte können die herkömmlichen, auf vollständigen Daten- oder Distanzmatrizen basierenden Auswertungsmethoden nicht mehr unmittelbar zur Anwendung kommen. Somit ergibt sich die Notwendigkeit einer expliziten Berücksichtigung der fehlenden Werte im Rahmen der datenanalytischen Untersuchung. In dieser Arbeit werden dabei vor allem die Möglichkeiten einer Analyse des Mechanismus, der den fehlenden Daten zugrundeliegt, sowie die darauf aufbauenden Verfahren zur Behandlung unvollständiger Daten- und Distanzmatrizen vorgestellt.

In der Literatur existieren mittlerweile zwar eine Vielzahl von Einzelarbeiten, die sich mit dem Problem fehlender Daten beschäftigen, eine grundlegende Gesamtdarstellung ist jedoch nicht zu finden. Diese Arbeit verfolgt daher das Ziel, sowohl die aus der Literatur bekannten wie auch die daraus ableitbaren bzw. darüber hinaus denkbaren Lösungsansätze und Methoden umfassend und mathematisch orientiert darzustellen und in ein Gesamtkonzept zur Auswertung einer unvollständigen Daten- bzw. Distanzmatrix zu integrieren.

Die Arbeit richtet sich somit an Wissenschaftler und Praktiker, die mit dem Problem unvollständiger Daten- und Distanzmatrizen konfrontiert werden. Für diesen Personenkreis soll die grundlegende Vorgehensweise zur adäquaten Behandlung fehlender Daten ausführlich dargelegt und erörtert werden.

Ich möchte dieses Vorwort nicht schließen, ohne mich bei all denjenigen sehr herzlich zu bedanken, die durch ihre konstruktiven Anregungen und ihre Diskussionsbereitschaft bei der Entstehung dieser Arbeit mitgewirkt haben. Zu großem Dank verpflichtet bin ich Herrn Prof. Dr. Otto Opitz und Herrn Prof. Dr. Günter Bamberg sowie allen Kolleginnen und Kollegen des Instituts für Statistik und Mathematische Wirtschaftstheorie der Universität Augsburg. Vor allem Herr Opitz hat wesentlich dazu beigetragen, daß die Arbeit in der vorliegenden Form zustande gekommen ist. Mein besonderer Dank gebührt darüber hinaus Herrn Dr. Wolfgang Hauke. Er hatte trotz der durchaus nervtötenden Tipparbeiten, die er aufgrund des gemeinsamen Büroraums gezwungenermaßen ertragen mußte, immer Zeit für meine Probleme.

Augsburg, im April 1995 Udo Bankhofer

Inhaltsverzeichnis

1 Einführung und Zielsetzung

Die Datenanalyse stellt sich die Aufgabe, die Ähnlichkeitsbeziehungen zwischen den Elementen einer endlichen **Objektmenge** $N = \{1,...,n\}$, die Teilmenge einer Grundgesamtheit von Objekten ist, zu analysieren. Den Ausgangspunkt für eine derartige Analyse bilden Daten, wobei unter anderem die folgenden drei Möglichkeiten einer vorliegenden **Datenbasis** bzw. **Datengrundlage** unterschieden werden können:

1. Die Ähnlichkeiten zwischen den Objekten werden indirekt durch **Merkmale**, die zur Charakterisierung der Objekte ausgewählt werden, zum Ausdruck gebracht.
2. Die Ähnlichkeiten zwischen den Objekten werden indirekt durch **Präferenzen**, die zur Beurteilung von Objektpaaren abgegeben werden, erfaßt.
3. Die Ähnlichkeiten zwischen den Objekten werden direkt durch **Ähnlichkeits-** bzw. **Verschiedenheitsmaße** quantifiziert.

Die jeweils vorliegenden Daten können in Form einer Matrix zusammengefaßt werden. Dabei sind in Abhängigkeit der drei vorgestellten Varianten einer denkbaren Datengrundlage die folgenden Arten von Matrizen zu unterscheiden:

1. **Datenmatrix** der Dimension $(n \times m)$, wobei m die Anzahl der zur Beschreibung der Objekte ausgewählten Merkmale bezeichnet
2. **Beurteilungsmatrix** der Dimension $(n \times n)$
3. **Distanzmatrix** der Dimension $(n \times n)$ im Fall der Verwendung von Verschiedenheitsmaßen

Während die Datenmatrix als Rohergebnis noch eine eher allgemeine Form besitzt, stellen die quadratische Beurteilungsmatrix und noch stärker die symmetrische Distanzmatrix speziellere Möglichkeiten zur Beschreibung der Ähnlichkeitsbeziehungen zwischen den Objekten dar.

Im Rahmen dieser Arbeit werden lediglich die Daten- sowie die Distanzmatrix als Ausgangspunkt einer Analyse der Ähnlichkeitsbeziehungen zwischen den Objekten betrachtet. Daher sollen diese beiden Formen einer Datengrundlage sowie die Beziehungen untereinander vor dem Hintergrund denkbarer fehlender Daten zunächst noch einmal ausführlicher dargestellt werden.

Erfolgt eine Beschreibung der Objekte durch ausgewählte Merkmale, dann können diese Merkmale in einer sogenannten **Merkmalsmenge** $M = \{1,...,m\}$ zusammengefaßt wer-

den. Falls bei einem oder mehreren Objekten nicht alle Ausprägungen bezüglich der erhobenen Merkmale vorliegen, erhält man eine **Datenmatrix** A der Form

$$A = (a_{ik})_{n,m} = \begin{pmatrix} a_{11} & \dots & \dots & a_{1m} \\ \vdots & \circ & & \vdots \\ & & & \circ \\ \vdots & & \circ & \vdots \\ a_{n1} & \dots & \dots & a_{nm} \end{pmatrix}. \tag{1.1}$$

Dabei entsprechen jeder Zeile die Ausprägungen eines Objektes bezüglich aller Merkmale und jeder Spalte die Ausprägungen eines Merkmals bei allen Objekten. Durch die Symbole ∘ werden in Anlehnung an *Toutenburg (1992, S. 197)* fehlende Ausprägungen angedeutet. Einige Methoden der Multivariaten Datenanalyse, wie die Faktoren-, Regressions- oder Diskriminanzanalyse, basieren unmittelbar auf der Datenmatrix, die jedoch im allgemeinen als vollständig vorausgesetzt wird. Andere Verfahren, wie beispielsweise die hierarchischen Klassifikationsverfahren oder die multidimensionale Skalierung, gehen von einer **Distanzmatrix** D der Form

$$D = (d_{ij})_{n,n} = \begin{pmatrix} d_{11} & \dots & d_{1n} \\ \vdots & & \vdots \\ d_{n1} & \dots & d_{nn} \end{pmatrix} \tag{1.2}$$

aus, wobei die Distanzindizes d_{ij} die Verschiedenheit von je zwei Objekten $i, j \in N$ quantifizieren. D kann entweder aus einer Datenmatrix abgeleitet oder aber direkt erhoben werden. Eine ausführliche Darstellung der Bestimmung merkmalsweiser Distanzindizes aus der Datenmatrix sowie deren Aggregation unter Berücksichtigung unterschiedlicher Skalentypen kann beispielsweise *Opitz (1980, S. 30-64)* entnommen werden. Grundsätzlich setzt eine derartige Berechnung jedoch eine vollständige Datenmatrix voraus. Die resultierende Distanzmatrix ist dann ebenfalls vollständig und kann problemlos in der weiteren Analyse verwendet werden. Im Vergleich dazu können bei der unmittelbaren Erhebung einer Distanzmatrix fehlende paarweise Distanzen auftreten. Eine auf diese Art ermittelte Distanzmatrix D besitzt dann die Form

$$D = (d_{ij})_{n,n} = \begin{pmatrix} d_{11} & \dots & \dots & d_{1n} \\ \vdots & \circ & & \vdots \\ & & & \circ \\ \vdots & & \circ & \vdots \\ d_{n1} & \dots & \dots & d_{nn} \end{pmatrix}, \tag{1.3}$$

wobei die Symbole ∘ die fehlenden paarweisen Distanzen andeuten. Da zwischen einer unmittelbar erhobenen Distanzmatrix und einer Datenmatrix kein direkter Zusammen-

hang besteht, sind auch die gegebenenfalls jeweils fehlenden Werte in keiner Beziehung zueinander.

Ausgangspunkt der nachfolgenden Betrachtungen ist damit entweder eine Datenmatrix der Form (1.1) oder eine unmittelbar erhobene Distanzmatrix der Form (1.3). Die Anwendung von herkömmlichen Methoden der Multivariaten Datenanalyse ist jedoch in beiden Fällen nicht möglich. Das Ziel dieser Arbeit besteht nun darin, die hier kurz skizzierte Problematik sowie deren Lösungsmöglichkeiten umfassend darzustellen, zu diskutieren und abzurunden. Dabei sollen sowohl die aus der Literatur bekannten wie auch die daraus ableitbaren bzw. darüber hinaus denkbaren Ansätze und Methoden in ein Gesamtkonzept zur Behandlung und Auswertung einer unvollständigen Daten- bzw. Distanzmatrix integriert werden.

Kapitel 2 widmet sich zunächst den Ursachen unvollständiger Daten- und Distanzmatrizen. Neben einer Darstellung allgemeiner Ursachen, die im Rahmen einer datenanalytischen Untersuchung zu fehlenden Werten führen können, wird vor allem zwischen systematischen und unsystematischen Ausfallmechanismen unterschieden. Während bei Vorliegen eines systematischen Ausfallmechanismus die Daten nicht zufällig fehlen, wird im Fall eines unsystematischen Ausfallmechanismus von einem zufälligen Fehlen der Daten gesprochen. Bei den unsystematischen Ausfallmechanismen sind darüber hinaus zwei Formen, die auch für Subgruppen von Objekten definiert werden können, gegeneinander abzugrenzen, und zwar das zufällige Fehlen innerhalb der an sich unbekannten Realisierungen der fehlenden Werte sowie innerhalb der gesamten Daten- bzw. Distanzmatrix.

Da lediglich bei Kenntnis der Art des Ausfallmechanismus eine problemadäquate Behandlung der fehlenden Daten existieren kann, werden in **Kapitel 3** Ansätze zur Strukturanalyse einer unvollständigen Daten- bzw. Distanzmatrix aufgezeigt. Grundsätzlich kann zwischen einer deskriptiven, einer explorativen sowie einer induktiven Strukturanalyse unterschieden werden. Die in Betracht zu ziehenden Ansätze und Methoden sollen dabei ausführlich dargestellt und hinsichtlich ihrer Aussagefähigkeit über den zugrundeliegenden Ausfallmechanismus beurteilt werden. Um die Grundlage für die weitere datenanalytische Untersuchung und damit für die Behandlung der fehlenden Daten vollständig bereitzustellen, sind die aus den Ergebnissen der Strukturanalyse resultierenden Konsequenzen zu diskutieren.

In **Kapitel 4** erfolgt schließlich eine Darstellung der Verfahren zur Behandlung fehlender Daten. Dabei lassen sich grundsätzlich die fünf Verfahrenskategorien Eliminierungsverfahren, Imputationsverfahren, Parameterschätzverfahren, multivariate Analyseverfahren und Sensitivitätsbetrachtungen unterscheiden. Während bei den Parameterschätzverfahren und den multivariaten Analyseverfahren eine unmittelbare Bestim-

mung von Verteilungsparametern bzw. multivariaten Analyseergebnissen aus dem unvollständigen Datenmaterial erfolgt, wird durch die Anwendung von Eliminierungs- und Imputationsverfahren eine vollständige und damit mit herkömmlichen Methoden auswertbare Datengrundlage bereitgestellt. Dabei werden im Fall einer Eliminierung Objekte bzw. Merkmale mit fehlenden Werten aus der Analyse ausgeschlossen und im Fall einer Imputation die fehlenden Daten durch geeignete Schätzwerte ersetzt. Eine eigene Verfahrenskategorie bilden schließlich noch die denkbaren Ansätze einer Sensitivitätsbetrachtung. Diese Ansätze verfolgen das Ziel, die Sensitivität von Imputationswerten, Parameterschätzungen oder Analyseergebnissen gegenüber dem zugrundegelegten Ausfallmechanismus und dem verwendeten Verfahren zur Behandlung der fehlenden Daten zu untersuchen. Bei der Darstellung der einzelnen Verfahren der fünf Kategorien sollen neben den jeweils benötigten Voraussetzungen bezüglich des Ausfallmechanismus, der Verteilung der Merkmalsausprägungen sowie des Skalenniveaus der Merkmale vor allem die jeweiligen Stärken und Schwächen sowie die Eignung im Rahmen einer datenanalytischen Untersuchung herausgearbeitet und kritisch gewürdigt werden.

Die in den Kapiteln 2, 3 und 4 vorgestellten Ansätze und Methoden sollen jeweils anhand von Beispielen erläutert werden. In Anhang A ist dazu eine Datenmatrix von 15 Statistiksoftwarepaketen, bei denen neun Merkmale erhoben wurden, angegeben. In Anhang B ist eine unmittelbar erhobene Distanzmatrix von 10 Mittelklasseautomobilen dargestellt. Bezüglich einer näheren Erläuterung der beiden Matrizen sei auf den entsprechenden Anhang verwiesen. Je nach Sachverhalt wird im folgenden, mit Ausnahme des Kapitels 2, nach dem Hinweis **„Beispiel:“** auf die Datenmatrix des Anhangs A oder auf die Distanzmatrix des Anhangs B oder auf beide Matrizen Bezug genommen, wobei jeweils eine explizite Nennung des verwendeten Datenmaterials erfolgt.

An dieser Stelle ist noch darauf hinzuweisen, daß aus Gründen einer möglichst einfachen Symbolik im Rahmen dieser Arbeit dasselbe Symbol für Zufallsvariablen und deren Realisierungen verwendet wird. Aus dem Zusammenhang wird jeweils ersichtlich sein, ob es sich um eine Zufallsvariable oder deren Realisierung handelt.

2 Ursachen unvollständiger Daten- und Distanzmatrizen

Die Frage nach den Ursachen unvollständiger Daten- und Distanzmatrizen ist prinzipiell mit der Frage nach dem Mechanismus, der zum Fehlen der Daten führt, verbunden. In der Literatur (vgl. z.B. *Schwab, 1991, S. 6*) spricht man in diesem Zusammenhang von **response mechanism**, **missing data mechanism**, **Antwort-** oder **Ausfallmechanismus**. Grundsätzlich sind die folgenden zwei Arten von fehlenden Daten, die häufig mit **MD** für **missing data** abgekürzt werden, zu unterscheiden (vgl. z.B. *Lösel, Wüstendörfer, 1974, S. 342*):

- **Unsystematisch** bzw. **zufällig fehlende Daten**
- **Systematisch** bzw. **nicht zufällig fehlende Daten**

Entsprechend wird auch zwischen unsystematischen bzw. vernachlässigbaren sowie systematischen bzw. nicht vernachlässigbaren Ausfallmechanismen unterschieden. Es stellt sich damit die Frage, wo und in welcher Art und Weise ein Ausfallmechanismus im Rahmen eines datenanalytischen Untersuchungsprozesses fehlende Werte erzeugen kann. Zur Lösung einer taxonomischen Aufgabenstellung sind im allgemeinen die folgenden fünf Stufen zu durchlaufen (vgl. z.B. *Opitz, 1980, S. 6-7*):

- **Festlegung von Untersuchungsgegenstand und Untersuchungsziel**
- **Diskussion der Datenbasis**
- **Datenerhebung**
- **Datenauswertung**
- **Interpretation der Ergebnisse**

Die Ursachen einer unvollständigen Daten- oder Distanzmatrix, die zur Auswertung vorliegt, sind folglich in den der Datenauswertung vorgelagerten Stufen des dargestellten Analyseablaufs zu suchen. Da nach der Festlegung von Untersuchungsgegenstand und Untersuchungsziel erst in Verbindung mit der Diskussion der Datenbasis, also der Charakterisierung der Objekte durch Auswahl von Merkmalen bzw. durch direkte Vergleiche, fehlende Daten hervorgerufen werden können, beschränkt sich die Suche nach den Ausfallursachen zunächst auf die beiden Stufen Diskussion der Datenbasis und Datenerhebung. Bei einer Auswertung mittels elektronischer Datenverarbeitung wird nach Abschluß einer Datenerhebung üblicherweise eine sogenannte Datenaufbereitung, die aus der Erfassung, Prüfung und Bereinigung der Daten besteht, durchgeführt. Die Datenaufbereitung kann als zusätzliche Stufe zwischen Datenerhebung und -auswertung

in dem oben dargestellten Untersuchungsablauf angesehen werden. Dabei ist die Verursachung fehlender Daten ebenfalls möglich.

In **Abschnitt 2.1** werden zunächst die bei der Diskussion der Datenbasis, der Datenerhebung und -aufbereitung denkbaren Ausfallursachen unabhängig vom zugrundeliegenden Ausfallmechanismus diskutiert. Da die Feststellung, ob die Daten zufällig oder systematisch fehlen, die weiteren Analyseschritte entscheidend beeinflußt, werden in **Abschnitt 2.2** unsystematische und in **Abschnitt 2.3** systematische Ausfallmechanismen vorgestellt. Der **Abschnitt 2.4** widmet sich schließlich speziellen Mustern fehlender Daten, die gerade in praktischen Anwendungen häufig auftreten, und der **Abschnitt 2.5** stellt die sich ergebenden Konsequenzen für die Analyse der fehlenden Daten dar. Zuvor sollen jedoch noch einige, für die Abschnitte 2.2 und 2.3 notwendige Vorüberlegungen durchgeführt werden.

Ausgangspunkt der nachfolgenden Betrachtungen ist entweder eine unvollständige Datenmatrix A der Form (1.1) oder eine unvollständige, unmittelbar erhobene Distanzmatrix D der Form (1.3). Es wird davon ausgegangen, daß die Möglichkeiten einer grundsätzlich denkbaren Nacherhebung von zunächst fehlenden Daten ausgeschöpft wurden, so daß die vorliegende Daten- bzw. Distanzmatrix nicht weiter vervollständigt werden kann.[1] Die tatsächlichen, an sich unbekannten Realisierungen der fehlenden Werte, die in (1.1) und (1.3) jeweils mit dem Symbol ∘ angedeutet sind, werden zusammenfassend mit A^{mis} bzw. D^{mis} und die jeweils vorhandenen Werte entsprechend mit A^{obs} bzw. D^{obs} bezeichnet [2], d.h.

$$A = \left(A^{obs}, A^{mis}\right), \tag{2.1}$$

$$D = \left(D^{obs}, D^{mis}\right). \tag{2.2}$$

Des weiteren wird für eine unvollständige Datenmatrix A eine **MD-Indikatormatrix** gemäß

$$V = \left(v_{ik}\right)_{n,m} = \begin{pmatrix} v_{11} & \cdots & v_{1m} \\ \vdots & & \vdots \\ v_{n1} & \cdots & v_{nm} \end{pmatrix} \text{ mit } v_{ik} = \begin{cases} 1 & \text{falls } a_{ik} \text{ vorhanden} \\ 0 & \text{sonst} \end{cases} \tag{2.3}$$

---

1 In realen Untersuchungen werden derartige Nacherhebungen aufgrund des damit verbundenen finanziellen Aufwands meist unterlassen oder sind sogar grundsätzlich nicht durchführbar. Dennoch stellt die Nacherhebung von zunächst fehlenden Daten die einfachste Möglichkeit dar, das Problem unvollständiger Daten- und Distanzmatrizen zu lösen.

2 A^{mis} und A^{obs} bzw. D^{mis} und D^{obs} bezeichnen jeweils lediglich die fehlenden und vorhandenen Realisierungen der Merkmalsausprägungen bzw. paarweisen Distanzen und stellen keine Matrizen im eigentlichen Sinn dar.

sowie für eine unvollständige, unmittelbar erhobene Distanzmatrix D eine **MD-Indikatormatrix** gemäß

$$W = (w_{ij})_{n,n} = \begin{pmatrix} w_{11} & \dots & w_{1n} \\ \vdots & & \vdots \\ w_{n1} & \dots & w_{nn} \end{pmatrix} \text{ mit } w_{ij} = \begin{cases} 1 & \text{falls } d_{ij} \text{ vorhanden} \\ 0 & \text{sonst} \end{cases} \tag{2.4}$$

eingeführt. Ein vollständiges Modell für die Daten und den Ausfallmechanismus wird schließlich durch die folgenden Verteilungen für die Daten- bzw. Distanzmatrix sowie deren Indikatormatrizen spezifiziert, wobei in diesem Fall A, V, D und W jeweils als Matrizen von Zufallsvariablen aufzufassen sind:

- A: Verteilung mit Parameter θ, Dichte- bzw. Wahrscheinlichkeitsfunktion $f(A \mid \theta)$
- V: Verteilung mit Parameter Θ, Wahrscheinlichkeitsfunktion $f(V \mid \Theta)$
- D: Verteilung mit Parameter ψ, Dichte- bzw. Wahrscheinlichkeitsfunktion $f(D \mid \psi)$
- W: Verteilung mit Parameter Ψ, Wahrscheinlichkeitsfunktion $f(W \mid \Psi)$

Beispiel:

Betrachtet man eine unvollständige Datenmatrix, in der die Erhebungsergebnisse von n Personen bezüglich der Merkmale Jahreseinkommen (1), Alter (2) und Geschlecht (3) zusammengefaßt sind, dann können beispielsweise die folgenden bedingten Wahrscheinlichkeiten bzw. Wahrscheinlichkeitsfunktionen für die Indikatormatrix V zugrundegelegt werden:

1. Für jede Person i (i = 1,...,n) sei die Wahrscheinlichkeit, daß die Ausprägung a_{ik} eines beliebigen Merkmals k (k = 1,2,3) jeweils vorhanden ist, gleich Θ ($0 \le \Theta \le 1$), d.h.

$$f(V \mid A, \Theta) = f(V \mid \Theta) = \prod_{i=1}^{n} \prod_{k=1}^{3} \Theta^{v_{ik}} \cdot (1-\Theta)^{1-v_{ik}} \; ; \; 0^0 = 1.$$

2. Für jede Person i (i = 1,...,n) sei die Wahrscheinlichkeit, daß die Ausprägung a_{i1} des Merkmals Jahreseinkommen vorhanden ist, gleich α ($0 \le \alpha \le 1$). Des weiteren sind für alle Personen die Ausprägungen a_{i3} des Merkmals Geschlecht mit der Wahrscheinlichkeit γ ($0 \le \gamma \le 1$) vorhanden, jedoch geben die weiblichen Personen, d.h. a_{i3} = weiblich, ihr Alter und damit die Ausprägungen a_{i2} mit der Wahrscheinlichkeit $1 - \beta$ ($0 \le \beta \le 1$) nicht an, während die männlichen Personen, d.h. a_{i3} = männlich, ihr Alter mit Sicherheit angeben. Damit ergibt sich die folgende bedingte Wahrscheinlichkeitsfunktion für V:

$$f(V \mid A, \alpha, \beta, \gamma) = \prod_{i=1}^{n} \alpha^{v_{i1}} \cdot (1-\alpha)^{1-v_{i1}} \cdot \left(1-(1-\beta) \cdot g(a_{i3})\right)^{v_{i2}} \cdot \left((1-\beta) \cdot g(a_{i3})\right)^{1-v_{i2}} \cdot \gamma^{v_{i3}} \cdot (1-\gamma)^{1-v_{i3}}$$

$$\text{mit } g(a_{i3}) = \begin{cases} 1 & \text{falls } a_{i3} = \text{weiblich} \\ 0 & \text{falls } a_{i3} = \text{männlich} \end{cases} ; \; 0^0 = 1.$$

3. Für jede Person i (i = 1,...,n) fehlt die Ausprägung a_{i1} des Merkmals Jahreseinkommen mit der Wahrscheinlichkeit $1 - \alpha$ ($0 \le \alpha \le 1$), wenn das Jahreseinkommen höher als 100000 DM liegt, während die Ausprägung bei einem Jahreseinkommen kleiner gleich 100000 DM mit Sicherheit vorhan-

den ist. Die Ausprägungen a_{i2} und a_{i3} der Merkmale Alter und Geschlecht seien bei allen Personen unabhängig von der realisierten Ausprägung mit Sicherheit vorhanden. Damit ergibt sich die folgende bedingte Wahrscheinlichkeitsfunktion für V:

$$f(V \mid A,\alpha) = \prod_{i=1}^{n} \left(1-(1-\alpha)\cdot h(a_{i1})\right)^{v_{i1}} \cdot \left((1-\alpha)\cdot h(a_{i1})\right)^{1-v_{i1}} \cdot 1^{v_{i2}} \cdot 0^{1-v_{i2}} \cdot 1^{v_{i3}} \cdot 0^{1-v_{i3}}$$

$$\text{mit } h(a_{i1}) = \begin{cases} 1 & \text{falls } a_{i1} > 100000 \\ 0 & \text{falls } a_{i1} \leq 100000 \end{cases}; \; 0^0 = 1.$$

2.1 Allgemeine Ausfallursachen

In Anlehnung an *Lösel und Wüstendörfer (1974, S. 343)* sowie *Schnell (1986, S. 24-56)* werden zunächst einige Ausfallursachen dargestellt, die unabhängig vom zugrundeliegenden Ausfallmechanismus von besonderer Bedeutung sind. Dabei erfolgt eine Einteilung nach den Ausfallursachen, die im Rahmen der Diskussion der Datenbasis, der Datenerhebung und der Datenaufbereitung auftreten können.

2.1.1 Diskussion der Datenbasis

Durch die Auswahl geeigneter Merkmale bzw. durch direkte Vergleiche kann eine Charakterisierung der Objekte erfolgen. Dabei ist das festgelegte Untersuchungsziel zu berücksichtigen. Man spricht in diesem Zusammenhang auch von der Festlegung eines Merkmalskatalogs bzw. allgemeiner von der Festlegung eines Untersuchungsdesigns. Generell sollte das Untersuchungsdesign so umfangreich wie nötig und so knapp wie möglich sein. Für das Auftreten von fehlenden Daten in der nachgelagerten Datenerhebung können unter anderem die folgenden Ursachen bei der Festlegung des Merkmalskatalogs bzw. des Untersuchungsdesigns genannt werden:

- **Fehlerhaftes Untersuchungsdesign:** Ein Untersuchungsdesign wird in der Art festgelegt, daß es mit Sicherheit bei der anschließenden Datenerhebung zu fehlenden Werten kommt. Als Beispiel sei das Merkmal „Alter der Kinder" genannt. Im Fall einer kinderlosen Person treten zwangsweise fehlende Werte auf, sofern diese Möglichkeit nicht durch eine entsprechende Gestaltung des Untersuchungsdesigns berücksichtigt wird.
- **Mangelhaftes Untersuchungsdesign:** Ein Untersuchungsdesign wird in der Art festgelegt, daß es bei der anschließenden Datenerhebung unter Umständen zu fehlenden Werten kommen kann. Hierzu zählen beispielsweise mißverständliche Fragen, unübersichtliche Anordnungen sowie selten verwendete Wörter und Redewendungen in Fragebögen.

2.1.2 Datenerhebung

Die Datenerhebung kann grundsätzlich nach den verwendeten Quellen in Sekundär- und Primäranalyse unterschieden werden. Bei einer Sekundäranalyse werden Daten verwendet, die zu einem anderen Zweck zu einem anderen Zeitpunkt erhoben wurden, aber im Prinzip denselben Sachverhalt im Hinblick auf das jetzt gestellte Untersuchungsziel beschreiben. Die folgenden Ausfallursachen sind in diesem Zusammenhang von besonderer Bedeutung:

- **Aktualitätsprobleme:** Die zur Verfügung stehenden Sekundärdaten sind unter Umständen teilweise veraltet. In diesem Fall sind lediglich die vorhandenen aktuellen Daten brauchbar und die veralteten Daten müssen als fehlende Werte betrachtet werden.
- **Verwendung mehrerer Sekundärquellen:** Bei der Verwendung mehrerer Sekundärquellen kann es vorkommen, daß die einzelnen Quellen Daten zu unterschiedlichen Objekten bereitstellen. Bei der Kombination dieser Quellen können dann einzelne Objekte fehlende Daten aufweisen.
- **Unvollständigkeit der Sekundärdaten:** Die Sekundärdaten können bereits unvollständig sein, so daß bei deren Verwendung die fehlenden Daten übernommen werden.

Im Rahmen einer Primäranalyse werden die Daten zur Erreichung des Untersuchungsziels explizit erhoben. Dabei werden zwei Erhebungstechniken unterschieden: Befragung und Beobachtung. Bei einer Befragung sind unter anderem die folgenden Ursachen für das Vorliegen fehlender Daten denkbar:

- **Übersehen von Fragen:** Bei einer schriftlichen Befragung übersieht der Befragte einzelne Fragen bzw. bei einer mündlichen Befragung werden einzelne Fragen vom Befrager übersehen und daher nicht gestellt.
- **Mangelndes Wissen:** Der Befragte kann trotz bester Bemühungen zu einzelnen Fragen keine Antwort geben. Dies kann vor allem bei Fragen zu länger zurückliegenden Ereignissen auftreten. Hierunter fallen nicht die sogenannten „Weiß nicht"-Antworten, die aufgrund der Unentschlossenheit oder Unentschiedenheit des Befragten gegebenen werden. Diese Antworten stellen keine fehlenden Daten, sondern Informationen dar.
- **Antwortverweigerung:** Der Befragte verweigert die Antwort zu einer schriftlich oder mündlich gestellten Frage. Dies kann vor allem bei Fragen auftreten, bei der eine Verletzung des Intimbereichs vermutet oder gesehen wird. So werden beispielsweise Fragen nach dem Einkommen, dem Trinkverhalten oder dem Sexualverhalten absichtlich nicht beantwortet. Des weiteren kann der Befragte Angst

vor den negativen Konsequenzen aus der Beantwortung einer Frage haben. Als Beispiel sei in diesem Zusammenhang die Frage nach dem Einkommen oder dem Vermögen genannt. Eine Antwortverweigerung kann darauf zurückzuführen sein, daß der Befragte steuerrechtliche Folgen im Fall der Kenntnisnahme durch das Finanzamt befürchtet.

- **Motivationsprobleme:** Die fehlenden Daten werden durch eine mangelnde Motivation des Befragten bzw. des Befragers hervorgerufen. Eine geringe Motivation des Befragten kann beispielsweise durch die Abneigung gegen die Befragungssituation, gegen den Interviewer oder gegen die Erhebung im allgemeinen sowie durch die Länge des Interviews hervorgerufen werden. Der Befrager kann z.B. aufgrund seiner Vergütung unmotiviert sein.
- **Verständnisprobleme:** Der Befragte hat Probleme die gestellte Frage inhaltlich zu verstehen und gibt daher keine oder eine unpassende Antwort.
- **Meinungslosigkeit:** Der Befragte hat zu einem Sachverhalt, den er beurteilen soll, keine Meinung oder kann seine Meinung nicht formulieren und gibt daher keine Antwort.
- **Zeitprobleme:** Der Befragte hat zur Beantwortung der Fragen eine vorher festgelegte Zeit und ist nicht in der Lage, innerhalb dieses Zeitlimits alle Fragen zu beantworten. Dadurch weisen meist die am Ende stehenden Fragen fehlende Werte auf.
- **Filterführung des Befragers:** Der Befrager überspringt absichtlich bestimmte Fragen. Dies kann beispielsweise dadurch begründet sein, daß der Befrager im Laufe der Befragung merkt, daß gewisse Fragen eine Reihe von Nachfragen bedingen. Durch Überspringen dieser Fragen kann der Befrager derartige Situationen vermeiden und die Befragung erheblich verkürzen.

Bei der Durchführung einer Beobachtung im Rahmen einer primärstatistischen Datenerhebung sind die folgenden Ausfallursachen von Bedeutung:

- **Unaufmerksamkeit des Beobachters:** Der Beobachter kann abgelenkt bzw. unaufmerksam sein und so den zu beobachtenden Sachverhalt verpassen.
- **Technische Probleme**: Wird die Beobachtung mittels Apparaten, wie beispielsweise Video- oder Blickaufzeichnung, durchgeführt, können aufgrund technischer Defekte die zu beobachtenden Sachverhalte verpaßt werden.
- **Motivationsprobleme:** Der Beobachter ist beispielsweise mit seiner Arbeitssituation oder seiner Vergütung unzufrieden und daher unmotiviert, die Beobachtung ordnungsgemäß durchzuführen.

Experimente, die zur Messung bzw. Aufdeckung der Auswirkungen von Veränderungen unabhängiger Einflußgrößen auf eine oder mehrere abhängige Größen dienen, stellen an sich keine Erhebungstechnik, sondern spezielle Versuchsanordnungen dar. Dabei können ebenfalls fehlende Daten hervorgerufen werden. Die folgenden Ursachen sollen beispielhaft genannt werden:

- **Zerstörung von Objekten:** Bestimmte Versuchsanordnungen, wie beispielsweise physikalische Experimente, können zur Zerstörung von Objekten führen.
- **Zensierung von Objekten:** Bei Studien über einen gewissen Zeitraum können einzelne Objekte durch Ausfall oder das Studienende zensiert sein. So kann beispielsweise bei klinischen Langzeitstudien, bei denen die Wirkung eines Medikaments untersucht werden soll, der Patient zwischenzeitlich eines natürlichen Todes sterben.

Bei einer Datenerhebung ist darüber hinaus zu unterscheiden, ob es sich um eine Vollerhebung oder eine Stichprobenerhebung handelt. Im Hinblick auf weitere Ausfallursachen spielt diese Unterscheidung jedoch keine Rolle, zumal im Rahmen einer datenanalytischen Untersuchung die in einer Stichprobe erfaßte Objektmenge ohnehin möglichst repräsentativ für die festgelegte Grundgesamtheit sein sollte (vgl. *Opitz, 1980, S. 7*).

2.1.3 Datenaufbereitung

Durch den Prozeß der Datenaufbereitung wird nach Abschluß der Datenerhebung aus dem Datenmaterial eine auswertbare Daten- oder Distanzmatrix erstellt. Im Fall einer Auswertung mittels elektronischer Datenverarbeitung erfolgt dabei in erster Linie die Erfassung, Prüfung und Bereinigung der Daten. Im Rahmen der Datenerfassung, bei der die erhobenen Daten üblicherweise codiert und auf einen maschinenlesbaren Datenträger übertragen werden, können unter anderem die folgenden Ursachen für das Vorliegen von fehlenden Werten genannt werden:

- **Codierfehler:** Bei der Codierung der Daten sind eine Reihe unterschiedlicher Fehler denkbar, die zum Vorliegen von fehlenden Daten führen. Beispielsweise können an sich vorhandene Daten irrtümlich als fehlend codiert werden. Auch ein beabsichtigtes Codieren vorhandener Daten als fehlend ist vorstellbar, falls es sich z.B. um unvorhergesehene Werte handelt. Des weiteren ist die Vergabe nicht vereinbarter Codes möglich, die für den Fall, daß die Ausgangsdaten später nicht mehr zugänglich sind, fehlende Daten darstellen.
- **Übertragungsfehler:** Bei der Übertragung der Daten auf einen maschinenlesbaren Datenträger sind sowohl elektronische wie auch manuelle Fehler vorstellbar.

Werden die Daten beispielsweise per Hand eingegeben, können Tippfehler vorkommen. Werden die Daten maschinell eingelesen, können Übertragungs- oder Speicherfehler zu fehlenden Daten führen.

Bei der Dateneditierung, die aus der Prüfung und der Bereinigung der Daten besteht, sind unter anderem die folgenden Ausfallursachen zu berücksichtigen:

- **Löschung von unmöglichen Daten:** Ein bei der Datenprüfung erkannter unmöglicher Wert wird aus dem Datenmaterial entfernt. So werden beispielsweise alphanumerische Zeichen bei einen Merkmal, das an sich nur numerische Ausprägungen besitzen kann, einfach gelöscht, ohne die in diesen Zeichen enthaltene Information zu nutzen bzw. deren Entstehen zu ergründen.
- **Löschung von fehlerhaften Daten:** Grundsätzlich realisierbare Daten werden bei der Datenprüfung als fehlerhaft erkannt bzw. angenommen und aus dem Datenmaterial entfernt.

2.2 Unsystematische Ausfallmechanismen

Zufällig bzw. unsystematisch fehlende Daten werden durch Einflußfaktoren auf eine Untersuchungssituation hervorgerufen, die sich nicht eindeutig auf bestimmte Objekte oder Merkmale konzentrieren. In diesem Zusammenhang sind vor allem Aufmerksamkeitsschwankungen bei der Datenerhebung oder der Datenaufbereitung zu nennen, die beispielsweise beim Beobachter, beim Befragten oder bei der datenaufbereitenden Person auftreten (vgl. *Lösel, Wüstendörfer, 1974, S. 343*).

In der Literatur existieren eine Reihe unterschiedlicher Definitionen für das Vorliegen eines unsystematischen Ausfallmechanismus. Ausgehend von einer unvollständigen Datenmatrix sind bei *Anderson et al. (1983, S. 416-417)* vier grundlegende Definitionen zusammengefaßt. Danach werden Daten als zufällig fehlend bezeichnet, wenn

- die fehlenden Daten über die Matrix gestreut sind, d.h. keine Konzentration von fehlenden Werten vorhanden ist,
- für ein Objekt eine fehlende Merkmalsausprägung unabhängig von jeder anderen Merkmalsausprägung ist,
- für ein gegebenes Merkmal die fehlenden Ausprägungen dieselbe Verteilung besitzen wie der gesamte Merkmalsvektor, d.h. kein Zusammenhang zwischen Ausfallmechanismus und Wertebereich eines Merkmals existiert,
- zwischen den fehlenden Daten zweier Merkmale kein Zusammenhang besteht.

Exakte Definitionen für einen unsystematischen Ausfallmechanismus gehen jedoch auf *Rubin (1976, S. 584-585)* zurück. Er hat in seiner Arbeit notwendige Bedingungen aufgestellt, um den Mechanismus, der die fehlenden Daten erzeugt, als vernachlässigbar bezeichnen zu können. Diese sollen im folgenden dargestellt und im Rahmen dieser Arbeit verwendet werden.

2.2.1 Missing at Random (MAR)

Ausgehend von einer unvollständigen Datenmatrix werden die Daten als **zufällig fehlend, missing at random** oder kurz **MAR** bezeichnet, wenn für jeden Wert von Θ die folgende Bedingung erfüllt ist:

$$f\left(V \mid \left(A^{obs}, A^{mis}\right), \Theta\right) = \text{konstant} \quad \forall\, A^{mis}. \tag{2.5}$$

Dies bedeutet, daß die Ausfallwahrscheinlichkeit bzw. das Fehlen der Ausprägungen unabhängig von den fehlenden Daten ist, aber von den vorhandenen Werten abhängen kann. Diese Definition läßt sich analog auf eine unvollständige, unmittelbar erhobene Distanzmatrix übertragen. Danach werden die paarweisen Distanzen als MAR bezeichnet, wenn für jeden Wert von Ψ die Bedingung

$$f\left(W \mid \left(D^{obs}, D^{mis}\right), \Psi\right) = \text{konstant} \quad \forall\, D^{mis} \tag{2.6}$$

erfüllt ist. Dies bedeutet, daß die Ausfallwahrscheinlichkeit bzw. das Fehlen der paarweisen Distanzen unabhängig von den fehlenden Daten ist, aber von den vorhandenen Werten abhängen kann.

Beispiel:

Den Ausgangspunkt stellt die im vorherigen Beispiel betrachtete unvollständige Datenmatrix mit den Erhebungsergebnissen von n Personen bezüglich der Merkmale Jahreseinkommen (1), Alter (2) und Geschlecht (3) dar. Für die fehlenden Daten, die den unter 1., 2. und 3. im vorherigen Beispiel dargestellten Wahrscheinlichkeitsfunktionen für V jeweils zugrunde liegen können, lassen sich die folgenden, hier entsprechend numerierten Aussagen über die Eigenschaft MAR treffen:

1. Unabhängig von A und Θ sind die fehlenden Daten immer MAR, da die Wahrscheinlichkeitsfunktion für V nicht von den Ausprägungen der Datenmatrix A und damit auch nicht von den an sich unbekannten Realisierungen der fehlenden Werte abhängt.
2. Die fehlenden Daten sind nur dann MAR, wenn entweder $v_{i3} = 1\ \forall\, i$ oder $\gamma = 1$ oder $\beta = 1$ gilt. Fehlen also Daten nur bei den Merkmalen Jahreseinkommen und Alter, dann fehlen diese unabhängig von den an sich unbekannten, tatsächlich realisierten Werten und hängen bestenfalls von den vorhandenen Ausprägungen des Merkmals Geschlecht ab. Dies ist der Fall, wenn alle Ausprägungen

bezüglich des Merkmals Geschlecht entweder unabhängig von der Ausfallwahrscheinlichkeit tatsächlich vorhanden sind ($v_{i3} = 1 \; \forall \; i$) oder aber a priori mit Sicherheit vorhanden sein müssen ($\gamma = 1$). Fehlen jedoch zusätzlich oder auch ausschließlich Werte beim Merkmal Geschlecht, dann fehlen die Daten nicht zufällig, da das Fehlen von den tatsächlichen Realisierungen der fehlenden Ausprägungen des Merkmals Geschlecht abhängig ist. Dieser Fall kann jedoch mit $\beta = 1$ nicht mehr eintreten, da dann unabhängig vom Geschlecht das Alter mit Sicherheit angegeben wird, und damit die noch verbleibenden fehlenden Daten wieder MAR sind.

3. Die fehlenden Daten sind nur dann MAR, wenn entweder $v_{i1} = 1 \; \forall \; i$ oder $\alpha = 1$ gilt. Die Daten fehlen also zufällig, wenn nur bei den Merkmalen Alter und Geschlecht und nicht beim Merkmal Jahreseinkommen Werte fehlen ($v_{i1} = 1 \; \forall \; i$). Sobald eine Person ihr Einkommen nicht angibt, fehlen die Daten nicht mehr zufällig, da das Fehlen von dem an sich unbekannten, tatsächlich realisierten Einkommen dieser Person abhängt. Im Fall $\alpha = 1$ sind die Ausprägungen des Merkmals Jahreseinkommen unabhängig von der Einkommenshöhe mit Sicherheit vorhanden, so daß dann die Eigenschaft MAR für die noch verbleibenden fehlenden Daten vorliegt.

Geht man von einer unvollständigen, unmittelbar erhobenen Distanzmatrix aus, dann sind die fehlenden Distanzen nicht MAR, wenn sie beispielsweise ab einer bestimmten Größe aufgrund von Datenerfassungsproblemen fehlen, und MAR, wenn sie unabhängig von den tatsächlich realisierten Werten fehlen.

2.2.2 Missing Completely at Random (MCAR)

Die vorhandenen Ausprägungen einer unvollständigen Datenmatrix werden als **zufällig beobachtet, observed at random** oder kurz **OAR** bezeichnet, wenn für jede Realisierung von A^{mis} und jeden Wert von Θ die nachfolgende Bedingung erfüllt ist:

$$f\left(V \,\middle|\, \left(A^{obs}, A^{mis}\right), \Theta\right) = \text{konstant} \quad \forall \; A^{obs} . \tag{2.7}$$

Verbal formuliert bedeutet dies, daß das Vorhandensein der Ausprägungen unabhängig von den beobachteten Werten ist. Im Fall einer unvollständigen, unmittelbar erhobenen Distanzmatrix werden die vorhandenen paarweisen Distanzen als OAR bezeichnet, falls für jede Realisierung von D^{mis} und jeden Wert von Ψ die Bedingung

$$f\left(W \,\middle|\, \left(D^{obs}, D^{mis}\right), \Psi\right) = \text{konstant} \quad \forall \; D^{obs} \tag{2.8}$$

erfüllt und damit das Vorhandensein der paarweisen Distanzen unabhängig von den beobachteten Werten ist.

Beispiel:

Betrachtet man wiederum die unvollständige Datenmatrix mit den Erhebungsergebnissen von n Personen bezüglich der Merkmale Jahreseinkommen (1), Alter (2) und Geschlecht (3), dann können für die vorhan-

denen Daten, analog zum Beispiel über die Eigenschaft MAR, die folgenden Aussagen über die Eigenschaft OAR getroffen werden:

1. Unabhängig von A und Θ sind die beobachteten Daten immer OAR, da die Wahrscheinlichkeitsfunktion für V nicht von den Ausprägungen der Datenmatrix A und damit auch nicht von den beobachteten Werten abhängt.

2. Die fehlenden Daten sind nur dann OAR, wenn entweder $v_{i3} = 0 \; \forall \; i$ oder $\gamma = 0$ oder $\beta = 1$ gilt. Sind also Daten nur bei den Merkmalen Jahreseinkommen und Alter vorhanden, dann sind diese unabhängig von den beobachteten Werten und hängen bestenfalls von den an sich unbekannten, tatsächlich realisierten Ausprägungen des Merkmals Geschlecht ab. Dies ist der Fall, wenn alle Ausprägungen des Merkmals Geschlecht entweder unabhängig von der Ausfallwahrscheinlichkeit tatsächlich fehlen ($v_{i3} = 0 \; \forall \; i$) oder aber a priori mit Sicherheit fehlen müssen ($\gamma = 0$). Sind jedoch zusätzlich oder auch ausschließlich Ausprägungen beim Merkmal Geschlecht vorhanden, dann sind die Daten nicht zufällig beobachtet, da das Vorhandensein von den beobachteten Ausprägungen des Merkmals Geschlecht abhängig ist. Dieser Fall kann jedoch mit $\beta = 1$ nicht mehr eintreten, da dann unabhängig vom Geschlecht das Alter mit Sicherheit angegeben wird und damit die beobachteten Daten wieder OAR sind.

3. Die fehlenden Daten sind nur dann OAR, wenn entweder $v_{i1} = 0 \; \forall \; i$ oder $\alpha = 1$ gilt. Die Daten sind also zufällig beobachtet, wenn nur bei den Merkmalen Alter und Geschlecht und nicht beim Merkmal Jahreseinkommen Werte vorhanden sind ($v_{i1} = 0 \; \forall \; i$). Sobald eine Person ihr Einkommen angibt, sind die Daten nicht mehr zufällig beobachtet, da das Vorhandensein von der Höhe des Einkommens dieser Person abhängt. Im Fall $\alpha = 1$ sind die Ausprägungen des Merkmals Jahreseinkommen unabhängig von der Einkommenshöhe mit Sicherheit vorhanden, so daß dann die Eigenschaft OAR für die vorhandenen Daten vorliegt.

Geht man von einer unvollständigen, unmittelbar erhobenen Distanzmatrix aus, dann sind die vorhandenen Distanzen nicht OAR, wenn sie beispielsweise nur bis zu einer bestimmten Größe aufgrund von Datenerfassungsproblemen vorhanden sind, und OAR, wenn sie unabhängig von den beobachteten Werten vorliegen.

Ausgehend von den Bedingungen MAR und OAR wird nun eine weitere Definition eingeführt. Danach sind die Ausprägungen bzw. paarweisen Distanzen genau dann **missing completely at random** oder kurz **MCAR**, wenn die Bedingungen MAR und OAR, also (2.5) und (2.7) bzw. (2.6) und (2.8) jeweils erfüllt sind, d.h.

$$f\left(V \mid \left(A^{obs}, A^{mis}\right), \Theta\right) = \text{konstant} \quad \forall \; A^{obs}, A^{mis}, \tag{2.9}$$

$$f\left(W \mid \left(D^{obs}, D^{mis}\right), \Psi\right) = \text{konstant} \quad \forall \; D^{obs}, D^{mis}. \tag{2.10}$$

Dies bedeutet schließlich, daß die fehlenden Daten zufällig fehlen und die vorhandenen Daten zufällig beobachtet sind, oder anders formuliert, das Fehlen der Ausprägungen

bzw. der paarweisen Distanzen steht in keiner Beziehung zu den fehlenden und den vorhandenen Werten der Daten- bzw. Distanzmatrix.

Beispiel:

Analog zu den Beispielen über die Eigenschaften MAR und OAR können, wiederum ausgehend von der unvollständigen Datenmatrix mit den Erhebungsergebnissen von n Personen bezüglich der Merkmale Jahreseinkommen (1), Alter (2) und Geschlecht (3), die folgenden Aussagen über die Eigenschaft MCAR getroffen werden:

1. Unabhängig von A und Θ sind die Daten immer MCAR, da die Wahrscheinlichkeitsfunktion für V nicht von den Ausprägungen der Datenmatrix A und damit weder von den an sich unbekannten Realisierungen der fehlenden Werte noch von den beobachteten Werten abhängt.
2. Die Daten sind nur dann MCAR, wenn $\beta = 1$ gilt. Nur in diesem Fall sind die Eigenschaften MAR und OAR gleichzeitig gegeben. Das Fehlen von Daten steht dann in keiner Beziehung zu den Merkmalen Jahreseinkommen, Alter und Geschlecht.
3. Die Daten sind nur dann MCAR, wenn $\alpha = 1$ gilt. Nur in diesem Fall sind die Eigenschaften MAR und OAR gleichzeitig gegeben. Das Fehlen von Daten steht dann in keiner Beziehung zu den Merkmalen Jahreseinkommen, Alter und Geschlecht.

Die Zusammenhänge zwischen den Eigenschaften MAR, OAR und MCAR sollen noch einmal anhand eines einfacheren Beispiels verdeutlicht werden. Ausgangspunkt ist jetzt eine Datenmatrix, in der die Erhebungsergebnisse von n Personen bezüglich der beiden Merkmale Einkommen und Alter zusammengefaßt sind, wobei lediglich beim Merkmal Einkommen fehlende Daten vorliegen. Damit können in Abhängigkeit von der Beziehung zwischen den fehlenden Daten und den beiden Merkmalen die folgenden Aussagen über das Vorliegen der Eigenschaften MAR, OAR und MCAR getroffen werden:

Das Fehlen der Daten bezüglich des Merkmals **Einkommen** ist vom		**Alter**	
		unabhängig	abhängig
Einkommen	unabhängig	MCAR	MAR, nicht OAR
	abhängig	OAR, nicht MAR	weder MAR, noch OAR

Besteht für das Fehlen einer Ausprägung beim Merkmal Einkommen für alle Objekte dieselbe Wahrscheinlichkeit unabhängig von Alter oder Einkommen, dann sind die Daten MCAR. Hängt die Wahrscheinlichkeit für das Fehlen einer Ausprägung beim Merkmal Einkommen nur vom Alter und nicht vom Einkommen ab, dann sind die Daten MAR aber nicht OAR. Ist die Wahrscheinlichkeit für das Fehlen einer Ausprägung beim Merkmal Einkommen nur vom Einkommen und nicht vom Alter abhängig, dann sind die Daten OAR aber nicht MAR. Hängt schließlich die Wahrscheinlichkeit für das Fehlen einer Ausprägung beim Merkmal Einkommen von Alter und Einkommen ab, dann sind die Daten weder MAR noch OAR. Berücksichtigt man zusätzlich die durchaus realistische Annahme, daß die Merkmale Alter und Einkommen positiv hoch korreliert sind, dann können die beiden Fälle, daß die Daten nur MAR bzw. nur OAR sind, nicht auftreten, d.h. die Daten können dann lediglich entweder MCAR oder weder MAR, noch OAR sein.

2.2.3 MAR und MCAR innerhalb von Klassen

Es ist denkbar, daß die Eigenschaften MAR, OAR und MCAR bei Betrachtung aller Objekte der Objektmenge nicht vorliegen, aber innerhalb bestimmter Subpopulationen, also bestimmter Mengen von Objekten bzw. Teilmengen der Objektmenge, erfüllt sind. Mit diesem Grundgedanken können weitere, durchaus realistischere Modelle für einen unsystematischen Ausfallmechanismus entwickelt werden (vgl. z.B. *Santos, 1981, S. 17, Schnell, 1986, S. 7-8*).

Um die Anwendung von speziellen Verfahren zur Behandlung der fehlenden Daten, die für diese Art von Ausfallmechanismen geeignet sind, zu ermöglichen, müssen die Subpopulationen das Ergebnis einer exhaustiven, disjunkten Klassifikation der Objektmenge darstellen. Damit sind diese Modelle allerdings nur bei Vorliegen einer unvollständigen Datenmatrix und nicht im Fall einer unmittelbar erhobenen unvollständigen Distanzmatrix geeignet. Diese Einschränkung ist darauf zurückzuführen, daß zwar alle Ausprägungen einer Datenmatrix, nicht aber alle paarweisen Distanzen einer Distanzmatrix eindeutig den Klassen einer disjunkten Klassifikation der Objekte zugeordnet werden können.

Ausgehend von einer unvollständigen Datenmatrix sowie einer exhaustiven, disjunkten Klassifikation $\mathcal{K}$ der Objektmenge N in s Klassen $K_1,\ldots,K_s$ werden die nicht vorhandenen Daten als **zufällig fehlend innerhalb der Klassen, missing at random within classes** oder kurz **MARC** bezeichnet, wenn für jeden Wert von Θ die Bedingung

$$f\left(V_{K_r} \middle| \left(A_{K_r}^{obs}, A_{K_r}^{mis}\right), \Theta\right) = \text{konstant} \quad \forall\, A_{K_r}^{mis} \quad (r = 1,\ldots,s) \tag{2.11}$$

erfüllt ist. Dabei bezeichnet der Index K_r den Teil der jeweiligen Matrix, dem die Objekte der Klasse r entsprechen. Die Bedingung (2.11) bedeutet, daß die Ausfallwahrscheinlichkeit bzw. das Fehlen der Ausprägungen innerhalb der Klassen unabhängig von den fehlenden Daten ist.

Die vorhandenen Daten werden als **zufällig beobachtet innerhalb der Klassen, observed at random within classes** oder kurz **OARC** bezeichnet, wenn für jede Realisierung von $A_{K_r}^{mis}$ und jeden Wert von Θ die Bedingung

$$f\left(V_{K_r} \middle| \left(A_{K_r}^{obs}, A_{K_r}^{mis}\right), \Theta\right) = \text{konstant} \quad \forall\, A_{K_r}^{obs} \quad (r = 1,\ldots,s) \tag{2.12}$$

erfüllt ist. Dabei bezeichnet der Index K_r wiederum den Teil der jeweiligen Matrix, dem die Objekte der Klasse r entsprechen. Die Bedingung (2.12) bedeutet, daß das Vorhan-

densein der Ausprägungen innerhalb der Klassen unabhängig von den beobachteten Werten ist. Sind die Bedingungen (2.11) und (2.12) gleichzeitig erfüllt, d.h.

$$f\left(V_{K_r} \,\middle|\, \left(A_{K_r}^{obs}, A_{K_r}^{mis}\right), \Theta\right) = \text{konstant} \quad \forall\, A_{K_r}^{obs}, A_{K_r}^{mis} \quad (r = 1,\ldots,s), \tag{2.13}$$

dann werden die Daten als **missing completely at random within classes** oder kurz **MCARC** bezeichnet. In diesem Fall fehlen die Daten innerhalb der Klassen zufällig und die vorhandenen Daten sind innerhalb der Klassen zufällig beobachtet, d.h. es besteht innerhalb der Klassen kein Zusammenhang zwischen dem Fehlen einer Ausprägung und den fehlenden sowie beobachteten Werten der Datenmatrix.

Beispiel:

Den Ausgangspunkt stellt die in den vorherigen Beispielen bereits betrachtete unvollständige Datenmatrix mit den Erhebungsergebnissen von n Personen bezüglich der Merkmale Jahreseinkommen (1), Alter (2) und Geschlecht (3) dar. Für die fehlenden Daten lassen sich dann analog die folgenden, wiederum entsprechend numerierten Aussagen über die Eigenschaften MARC, OARC und MCARC treffen, wobei die für die Klassen relevanten Wahrscheinlichkeitsfunktionen jeweils aus der Wahrscheinlichkeitsfunktion für V abgeleitet und angegeben werden:

1. Die aus der Wahrscheinlichkeitsfunktion für V abgeleiteten Wahrscheinlichkeitsfunktionen innerhalb der s Klassen ergeben sich in allgemeiner Form gemäß

$$f\left(V_{K_r} \,\middle|\, A_{K_r}, \Theta\right) = f\left(V_{K_r} \,\middle|\, \Theta\right) = \prod_{i \in K_r} \prod_{k=1}^{3} \Theta^{v_{ik}} \cdot (1-\Theta)^{1-v_{ik}} \quad (r = 1,\ldots,s)\,;\; 0^0 = 1.$$

   Unabhängig von der vorgenommenen exhaustiven, disjunkten Klassifikation in s Klassen sind die Daten innerhalb der Klassen immer MAR, OAR und MCAR, d.h. die Eigenschaften MARC, OARC und MCARC sind erfüllt. Diese Tatsache ist auch aufgrund der bereits getroffenen Feststellungen, daß die Daten auch unklassifiziert die drei Eigenschaften erfüllen, ersichtlich.

2. Die aus der Wahrscheinlichkeitsfunktion für V abgeleiteten Wahrscheinlichkeitsfunktionen innerhalb der s Klassen ergeben sich in allgemeiner Form gemäß

$$f\left(V_{K_r} \,\middle|\, A_{K_r}, \alpha, \beta, \gamma\right) = \prod_{i \in K_r} \alpha^{v_{i1}} \cdot (1-\alpha)^{1-v_{i1}} \cdot \left(1-(1-\beta)\cdot g(a_{i3})\right)^{v_{i2}} \cdot \left((1-\beta)\cdot g(a_{i3})\right)^{1-v_{i2}} \cdot \gamma^{v_{i3}} \cdot (1-\gamma)^{1-v_{i3}}$$

$$(r = 1,\ldots,s) \;\text{ mit }\; g(a_{i3}) = \begin{cases} 1 & \text{falls } a_{i3} = \text{weiblich} \\ 0 & \text{falls } a_{i3} = \text{männlich} \end{cases};\; 0^0 = 1.$$

   Legt man nun beispielsweise eine exhaustive, disjunkte Klassifikation in der Art zugrunde, daß die Klasse K_1 alle männlichen Personen und die Klasse K_2 alle weiblichen Personen enthält, dann ergeben sich die folgenden beiden speziellen Wahrscheinlichkeitsfunktionen für V_{K_1} und V_{K_2}:

$$f\left(V_{K_1} \,\middle|\, A_{K_1}, \alpha, \beta, \gamma\right) = f\left(V_{K_1} \,\middle|\, \alpha, \beta, \gamma\right) = \prod_{i \in K_1} \alpha^{v_{i1}} \cdot (1-\alpha)^{1-v_{i1}} \cdot 1^{v_{i2}} \cdot 0^{1-v_{i2}} \cdot \gamma^{v_{i3}} \cdot (1-\gamma)^{1-v_{i3}},$$

$$f\left(V_{K_2} \,\middle|\, A_{K_2}, \alpha, \beta, \gamma\right) = f\left(V_{K_2} \,\middle|\, \alpha, \beta, \gamma\right) = \prod_{i \in K_2} \alpha^{v_{i1}} \cdot (1-\alpha)^{1-v_{i1}} \cdot \beta^{v_{i2}} \cdot (1-\beta)^{1-v_{i2}} \cdot \gamma^{v_{i3}} \cdot (1-\gamma)^{1-v_{i3}}.$$

Damit sind die Daten innerhalb der beiden Klassen immer MCAR, d.h. die nicht vorhandenen Daten fehlen innerhalb der Klassen zufällig, sind also MARC, die vorhandenen Daten sind innerhalb der Klassen zufällig beobachtet, sind also OARC, und insgesamt sind die Daten MCARC. Jede andere Klassifikation der Objekte führt in diesem Beispiel hinsichtlich der drei Eigenschaften dann zu einem identischen Ergebnis, wenn in einer Klasse keine männlichen und weiblichen Personen gleichzeitig sind.

3. Die aus der Wahrscheinlichkeitsfunktion für V abgeleiteten Wahrscheinlichkeitsfunktionen innerhalb der s Klassen ergeben sich in allgemeiner Form gemäß

$$f\left(V_{K_r} \mid A_{K_r}, \alpha\right) = \prod_{i \in K_r} \left(1-(1-\alpha)\cdot h(a_{i1})\right)^{v_{i1}} \cdot \left((1-\alpha)\cdot h(a_{i1})\right)^{1-v_{i1}} \cdot 1^{v_{i2}} \cdot 0^{1-v_{i2}} \cdot 1^{v_{i3}} \cdot 0^{1-v_{i3}}$$

$$(r = 1,\ldots,s) \text{ mit } h(a_{i1}) = \begin{cases} 1 & \text{falls } a_{i1} > 100000 \\ 0 & \text{falls } a_{i1} \leq 100000 \end{cases}; \; 0^0 = 1.$$

Ausgehend von einer beispielhaft festgelegten exhaustiven und disjunkten Klassifikation mit drei Klassen, wobei K_1 alle Personen mit einem Jahreseinkommen unter 50000 DM, K_2 alle Personen mit einem Jahreseinkommen von 50000 DM bis einschließlich 100000 DM sowie K_3 alle Personen mit einem Jahreseinkommen über 100000 DM enthält, ergeben sich die folgenden speziellen Wahrscheinlichkeitsfunktionen für V_{K_1}, V_{K_2} und V_{K_3}:

$$f\left(V_{K_1} \mid A_{K_1}, \alpha\right) = f\left(V_{K_1}\right) = \prod_{i \in K_1} 1^{v_{i1}} \cdot 0^{1-v_{i1}} \cdot 1^{v_{i2}} \cdot 0^{1-v_{i2}} \cdot 1^{v_{i3}} \cdot 0^{1-v_{i3}},$$

$$f\left(V_{K_2} \mid A_{K_2}, \alpha\right) = f\left(V_{K_2}\right) = \prod_{i \in K_2} 1^{v_{i1}} \cdot 0^{1-v_{i1}} \cdot 1^{v_{i2}} \cdot 0^{1-v_{i2}} \cdot 1^{v_{i3}} \cdot 0^{1-v_{i3}},$$

$$f\left(V_{K_3} \mid A_{K_3}, \alpha\right) = f\left(V_{K_3} \mid \alpha\right) = \prod_{i \in K_3} \alpha^{v_{i1}} \cdot (1-\alpha)^{1-v_{i1}} \cdot 1^{v_{i2}} \cdot 0^{1-v_{i2}} \cdot 1^{v_{i3}} \cdot 0^{1-v_{i3}}.$$

In diesem Fall sind die Daten innerhalb der drei Klassen immer MCAR, d.h. die Eigenschaft MCARC ist erfüllt. Damit fehlen die nicht vorhandenen Daten innerhalb der Klassen zufällig, sind also MARC, und die vorhandenen Daten sind innerhalb der Klassen zufällig beobachtet, sind also OARC. Jede andere Klassifikation der Objekte führt in diesem Beispiel hinsichtlich der drei Eigenschaften dann zu einem identischen Ergebnis, wenn in einer Klasse Personen mit einem Jahreseinkommen kleiner gleich 100000 DM sowie Personen mit einem Jahreseinkommen über 100000 DM nicht gleichzeitig sind.

Es stellt sich nun die Frage, inwieweit ein Zusammenhang zwischen einem unsystematischen Ausfallmechanismus bezogen auf die gesamte Datenmatrix A und bezogen auf die Teile der Datenmatrix A_{K_r}, die den Objekten der Klasse r entsprechen, existiert. Diese Frage stellt sich vor allem deshalb, da durch eine Klassifikation der Objekte weitere Informationen zur Verfügung gestellt werden, die für einen derartigen Zusammenhang genützt werden können. Das nachfolgend dargestellte und bewiesene Theorem bringt diese grundsätzlichen Überlegungen zum Ausdruck.

Theorem:

Wenn die fehlenden Daten einer unvollständigen Datenmatrix der Eigenschaft MAR genügen und eine Klassifikation der Objekte auf Basis der beobachteten Daten in der Art vorliegt, daß die Objekte einer Klasse identische Ausprägungen bezüglich der jeweils vorhandenen Werte besitzen, dann sind die Daten innerhalb der auf diese Weise definierten Klassen MCAR, d.h. die Eigenschaft MCARC gilt.

Beweis:

Aus der vorausgesetzten Bedingung MAR für die fehlenden Daten der unvollständigen Datenmatrix folgt, daß die fehlenden Daten auch innerhalb einzelner Klassen der Bedingung MAR, also insgesamt der Bedingung MARC, genügen, d.h. für jeden Wert von Θ gilt die Implikation

$$f\left(V \mid \left(A^{obs}, A^{mis}\right), \Theta\right) = \text{konstant} \quad \forall\, A^{mis}$$

$$\Rightarrow \quad f\left(V_{K_r} \mid \left(A_{K_r}^{obs}, A_{K_r}^{mis}\right), \Theta\right) = \text{konstant} \quad \forall\, A_{K_r}^{mis} \quad (r = 1,\ldots,s).$$

Wird zusätzlich eine Klassifikation der Objekte auf Basis aller beobachteten Daten in der Art vorausgesetzt, daß die Objekte einer Klasse identische Ausprägungen bezüglich der jeweils vorhandenen Werte besitzen, d.h.

$$\exists\, \mathcal{K} = \{K_1,\ldots,K_s\} \text{ mit } K_r = \left\{ i \in N: \quad (v_{ik} = 1) \Rightarrow (a_{ik} = c_{rk}),\, k \in M \right\} \quad (r = 1,\ldots,s),$$

wobei c_{rk} die Ausprägung der Klasse r bezüglich des Merkmals k bezeichnet, dann erfüllen auch die beobachteten Daten innerhalb der Klassen die Bedingung OAR, also insgesamt die Bedingung OARC, d.h.

$$f\left(V_{K_r} \mid \left(A_{K_r}^{obs}, A_{K_r}^{mis}\right), \Theta\right) = \text{konstant} \quad \forall\, A_{K_r}^{obs} \quad (r = 1,\ldots,s),$$

da damit nur eine Realisierung von $A_{K_r}^{obs}$ in Form der Klassenausprägungen bezüglich der einzelnen Merkmale möglich ist. Mit MARC und OARC gilt schließlich aufgrund der definierten Äquivalenz die Bedingung MCARC, d.h.

$$f\left(V_{K_r} \mid \left(A_{K_r}^{obs}, A_{K_r}^{mis}\right), \Theta\right) = \text{konstant}\ \forall\, A_{K_r}^{mis} \wedge f\left(V_{K_r} \mid \left(A_{K_r}^{obs}, A_{K_r}^{mis}\right), \Theta\right) = \text{konstant}\ \forall\, A_{K_r}^{obs}$$

$$\Leftrightarrow \quad f\left(V_{K_r} \mid \left(A_{K_r}^{obs}, A_{K_r}^{mis}\right), \Theta\right) = \text{konstant}\ \forall\, A_{K_r}^{obs}, A_{K_r}^{mis} \quad (r = 1,\ldots,s). \qquad \blacksquare$$

Damit kann also für den Fall, daß die vorliegenden Daten nur der Eigenschaft MAR und nicht der Eigenschaft MCAR genügen, durch Verwendung einer wie in obigem Theorem beschriebenen Klassifikation der Objekte die Eigenschaft MCAR innerhalb dieser Klas-

sen gefolgert werden. Jedoch ist die Bedingung, daß die Objekte einer Klasse identische Ausprägungen bezüglich der jeweils vorhandenen Werte besitzen, für die praktische Anwendung des Theorems zu restriktiv, da eine derartige Klassifikation der Objekte im allgemeinen zu Klassen mit einer sehr geringen Anzahl von Objekten oder sogar einelementigen Klassen führt. Es bietet sich daher an, eine abgeschwächte Bedingung für die zugrundegelegte Klassifikation zu formulieren, die dann allerdings auch nur eine eingeschränkte Folgerung in Form einer Approximation über die Bedingung MCAR innerhalb dieser Klassen zuläßt. Gemäß diesen Überlegungen wird von einer Klassifikation $\mathcal{K}$ der Objekte auf Basis der beobachteten Daten in der Art ausgegangen, daß bei gegebener Klassenanzahl s die Ähnlichkeit der Objekte einer Klasse K_r ($r = 1,...,s$) bezüglich der jeweils vorhandenen Werte maximal sein soll. Dies bedeutet, daß die Summe der paarweisen Distanzen von Objekten einer Klasse über alle Klassen minimal ist, d.h.

$$\sum_{K_r \in K} \sum_{i,j \in K_r} d_{ij}^{obs} \quad \rightarrow . \ \min .$$

Dabei zeigt der Index *obs* bei d_{ij}^{obs} an, daß es sich um Distanzen auf Basis der beobachteten Daten handelt. Im Idealfall ist diese Summe gleich Null, und zwar dann, wenn die Objekte einer Klasse K_r ($r = 1,...,s$) identische Ausprägungen bezüglich der jeweils vorhandenen Werte besitzen. In diesem Fall gilt dann nach obigem Theorem, daß mit der zusätzlich vorausgesetzten Bedingung MAR für die fehlenden Daten die Daten insgesamt der Bedingung MCARC genügen. Je ähnlicher die Objekte einer Klasse sind, desto eher wird die Bedingung MCARC erfüllt sein. Falls also eine Klassifikation auf Basis der beobachteten Daten in der Art vorliegt, daß bei gegebener Klassenanzahl die Ähnlichkeit der Objekte in den einzelnen Klassen bezüglich der jeweils vorhandenen Werte maximal wird, dann ist die Wahrscheinlichkeit, daß die Daten innerhalb der auf diese Weise definierten Klassen MCAR sind und damit insgesamt der Eigenschaft MCARC genügen, ebenfalls maximal.

Die Bedeutung dieser Ergebnisse wird bei der im nächsten Kapitel noch folgenden Darstellung der Konsequenzen des vorliegenden Ausfallmechanismus für die Behandlung der fehlenden Daten ersichtlich werden.

2.3 Systematische Ausfallmechanismen

Zur Feststellung, ob der zu den fehlenden Daten führende Mechanismus systematisch oder unsystematisch ist, muß ausschließlich die vorliegende unvollständige Daten- oder Distanzmatrix herangezogen werden. Es erfolgt also lediglich eine Betrachtung, inwieweit der datengenerierende Prozeß von den fehlenden bzw. beobachteten Daten abhängt und ob damit die vorhandenen Daten eine zufällige Stichprobe der Grundgesamtheit

darstellen. Entsprechend spielt es keine Rolle, inwiefern zum einen die Abhängigkeit zwischen dem Ausfallmechanismus und dem vorliegenden Datenmaterial tatsächlich gegeben und zum anderen der Ausfallmechanismus eventuell von nicht vorliegenden Daten, d.h. Daten die für die Untersuchung nicht von Bedeutung sind und daher nicht erhoben wurden, abhängig ist.

Beispiel:

Im Fall einer zeitlichen Begrenzung für das Ausfüllen eines Fragebogens werden unter Umständen die am Ende stehenden Fragen von einigen Personen, die zum Ausfüllen des Fragebogens mehr Zeit benötigen, nicht mehr beantwortet. Der Ausfallmechanismus ist damit zwar a priori unabhängig vom Datenmaterial, jedoch können die vorliegenden Daten von den fehlenden oder den beobachteten Werten in dem Sinn abhängig sein, daß sie keine zufällige Stichprobe der vollständigen Daten darstellen.

Damit kann also auch ein vom vorliegenden Datenmaterial unabhängiger Ausfallmechanismus anhand der in den Abschnitten 2.2.1 bis 2.2.3 aufgestellten Bedingungen, die sich prinzipiell nur auf die im Rahmen der Untersuchung interessierenden Daten beziehen, als vernachlässigbar bzw. nicht vernachlässigbar eingestuft werden.

Für die in Abschnitt 2.1 diskutierten allgemeinen Ausfallursachen können zum größten Teil sehr einfach systematische Mechanismen unterstellt werden. Für das Vorliegen systematischer Ausfallmechanismen ist jedoch nicht von Bedeutung, ob die zu den fehlenden Daten führenden Ursachen selbst systematischer Natur sind. Entscheidend sind lediglich die Auswirkungen auf das vorliegende Datenmaterial. So kann selbst das an sich zufällige Fehlen von Ausprägungen eines Merkmals mit einer bestimmten Wahrscheinlichkeit einen systematischen Mechanismus darstellen, wenn damit die Daten für eine Subgruppe fehlen. Im Gegensatz dazu hat beispielsweise das an sich systematisch erscheinende Fehlen von Daten aufgrund einer Filterführung des Befragers bei bestimmten Personen einen unsystematischen Ausfallmechanismus zur Folge, falls die letztendlich vorliegenden Daten eine zufällige Stichprobe darstellen.

Ein **systematischer Ausfallmechanismus** liegt grundsätzlich dann vor, wenn die fehlenden Daten nicht der Eigenschaft MAR genügen. Damit ist auch unabhängig vom Vorliegen der Eigenschaft OAR die Bedingung MCAR nicht erfüllt. Diese Aussage kann analog bei Vorliegen einer exhaustiven, disjunkten Klassifikation der Objekte formuliert werden. Ein **systematischer Ausfallmechanismus innerhalb der Klassen** ist also dann gegeben, wenn die fehlenden Daten innerhalb der Klassen nicht der Eigenschaft MAR, also die fehlenden Daten insgesamt damit nicht der Eigenschaft MARC genügen.

Beispiel:

Betrachtet man wiederum die unvollständige Datenmatrix mit den Erhebungsergebnissen von n Personen bezüglich der Merkmale Jahreseinkommen (1), Alter (2) und Geschlecht (3), dann können für die Daten

die folgenden, wiederum entsprechend numerierten Aussagen über das Vorliegen systematischer Ausfallmechanismen gemacht werden:

1. Grundsätzlich liegt kein systematischer Ausfallmechanismus vor, da die fehlenden Daten immer MAR sind.

2. Ein systematischer Ausfallmechanismus liegt dann vor, wenn $v_{i3} \neq 1$ für mindestens ein i, $\gamma \neq 1$ und $\beta \neq 1$ gilt. In diesem Fall ist die Eigenschaft MAR nicht erfüllt. Die vorhandenen Daten beim Merkmal Alter stellen dann keine zufällige Stichprobe der Grundgesamtheit dar, da die weiblichen Personen in der Stichprobe zu wenig berücksichtigt werden.

3. Ein systematischer Ausfallmechanismus liegt dann vor, wenn $v_{i1} \neq 1$ für mindestens ein i und $\alpha \neq 1$ gilt. In diesem Fall ist die Eigenschaft MAR nicht erfüllt. Die vorhandenen Daten stellen dann keine zufällige Stichprobe der Grundgesamtheit dar, da die Personen mit einem Jahreseinkommen über 100000 DM in der Stichprobe zu wenig berücksichtigt werden.

2.4 Spezielle Muster fehlender Daten

Wie in diesem Kapitel bereits deutlich wurde, können fehlende Daten im Laufe einer datenanalytischen Untersuchung aufgrund sehr unterschiedlicher Ursachen entstehen. Es gibt jedoch eine Reihe von speziellen Mustern fehlender Daten, die gerade in praktischen Anwendungen häufig auftreten. Im folgenden sollen einige dieser speziellen Muster fehlender Daten, deren zugrundeliegende Ausfallmechanismen größtenteils systematischer Art sind, vorgestellt werden.

Im Fall des Vorliegens einer unvollständigen Datenmatrix verwenden *Hill und Dixon (1981, S. 57)* die Bezeichnungen **truncated** und **related** für zwei spezielle Muster fehlender Daten. Ausgehend von einem kardinal skalierten Merkmal liegt ein sogenanntes **truncated MD-Muster** dann vor, wenn eine Ausprägung, die größer als das arithmetische Mittel dieses Merkmals ist, mit einer bestimmten Wahrscheinlichkeit fehlt. In diesem Fall ist die Annahme der Eigenschaft MAR für die Daten nicht mehr gerechtfertigt. Ein sogenanntes **related MD-Muster** liegt vor, wenn für ein Objekt $i \in N$, dessen Ausprägung beim Merkmal $k \in M$ größer als das arithmetische Mittel dieses Merkmals ist, die Ausprägung beim Merkmal $l \in M$, $l \neq k$ mit einer bestimmten Wahrscheinlichkeit fehlt. Die Eigenschaft OAR für die Daten ist in diesem Fall nicht erfüllt. Die bislang lediglich für kardinale Merkmale gemachten Ausführungen können analog auf ordinale Merkmale übertragen werden, wobei anstelle des arithmetischen Mittels der Median den geeigneten Lageparameter darstellt. Bei Vorliegen einer unvollständigen, unmittelbar erhobenen Distanzmatrix ist das MD-Muster im Hinblick auf den beschriebenen Ansatz dann truncated, wenn für ein Objekt $i \in N$ die paarweise Distanz zu einem anderen Objekt $j \in N, j \neq i$, die größer als das Mittel aller paarweisen Distanzen zwischen dem

Objekt $i \in N$ und den anderen Objekten $j \in N, j \neq i$ ist, mit einer bestimmten Wahrscheinlichkeit fehlt. Hingegen ist der Fall eines related MD-Musters bei Vorliegen einer Distanzmatrix nicht möglich, da das Fehlen einer paarweisen Distanz nicht von dem Wert einer anderen paarweisen Distanz abhängen kann.

Zwei zum truncated MD-Muster ähnliche, spezielle Muster fehlender Daten werden von *Nordheim (1978a, S. 58, 1978b, S. 36)* mit **probit missingness** und **logistic missingness** bezeichnet. Ausgehend von einer unvollständigen Datenmatrix sowie kardinalen Merkmalen bezeichnet probit missingness den Fall, daß die Ausprägung a_{ik} des Merkmals $k \in M$ bei Objekt $i \in N$ mit der Wahrscheinlichkeit

$$P(a_{ik}) = \Phi\left(\frac{a_{ik} - c}{t}\right) \quad (c, t \in \mathbb{R}; t > 0) \tag{2.14}$$

fehlt, wobei Φ die Verteilungsfunktion der Standardnormalverteilung ist. Durch den Parameter c kann eine Art Schwellenwert festgelegt werden, ab dem das Fehlen einer Ausprägung wahrscheinlicher ist als das Vorhandensein, während der Parameter t das Ausmaß steuert, mit dem der Wert der Ausprägung die Wahrscheinlichkeit für das Fehlen beeinflußt. Für sehr große Werte von t hängt damit die Wahrscheinlichkeit für das Fehlen einer Ausprägung nicht mehr von deren Wert ab, so daß in diesem Fall die Annahme der Eigenschaft MAR für die Daten gerechtfertigt ist. Fehlt die Ausprägung a_{ik} des Merkmals $k \in M$ bei Objekt $i \in N$ mit der Wahrscheinlichkeit

$$P(a_{ik}) = \frac{1}{1 + e^{a + b \cdot a_{ik}}} \quad (a, b \in \mathbb{R}) \tag{2.15}$$

dann wird dies mit logistic missingness bezeichnet. Mit den Parametern a und b wird wiederum festgelegt, wie der Wert der Ausprägung die Wahrscheinlichkeit für das Fehlen beeinflußt. Für $b = 0$ ist die Wahrscheinlichkeit für das Fehlen einer Ausprägung vom Wert der Ausprägung unabhängig, so daß damit die Eigenschaft MAR für die Daten vorliegt.

Ein weiteres spezielles Muster fehlender Daten stellt der Fall eines **vollständig fehlenden Objektvektors** dar, d.h. ausgehend von einer Datenmatrix fehlen sämtliche Ausprägungen bei einem Objekt $i \in N$ bzw. ausgehend von einer unmittelbar erhobenen Distanzmatrix fehlen sämtliche paarweise Distanzen zwischen einem Objekt $i \in N$ und den anderen Objekten $j \in N, j \neq i$. In der Literatur (vgl. z.B. *Schwab, 1991, S. 5*) wird dies mit **unit nonresponse** im Gegensatz zu **item nonresponse**, bei dem für ein Objekt nur einzelne Daten fehlen, bezeichnet. Ein vollständig fehlender Objektvektor kann sowohl auf einen systematischen wie auch auf einen unsystematischen Ausfallmechanismus zurückzuführen sein. Ein zufälliges Fehlen liegt vermutlich dann vor, wenn die Daten für ein Objekt aufgrund von Übertragungsfehlern in einer EDV-Anlage verloren

gehen. Ein fehlender Objektvektor bei einer Befragung zu den üblichen Geschäftszeiten deutet hingegen auf ein systematischen Fehlen hin, wenn die zu befragende Person berufstätig ist und deshalb nicht zu Hause angetroffen wird. Im Fall einer unvollständigen Datenmatrix ist auch das Vorliegen eines **vollständig fehlenden Merkmalsvektors** als spezielles Muster fehlender Daten denkbar. Dieses Muster spricht jedoch stark für ein systematisches Fehlen der Daten, da das zufällige Fehlen der Ausprägungen a_{ik} eines Merkmals k bei allen Objekten $i = 1,...,n$ für hinreichend große n eher einen Ausnahmefall darstellt.

Abschließend sollen noch kurz zwei spezielle MD-Muster vorgestellt werden, die im Vergleich zum Vorliegen eines vollständig fehlenden Objekt- bzw. Merkmalsvektors eine Abschwächung darstellen und in realen datenanalytischen Untersuchungen häufig anzutreffen sind. Ausgehend von einer unvollständigen Datenmatrix können zum einen die Daten für ein einzelnes Objekt oder für wenige Objekte größtenteils fehlen. Dieser Fall deutet darauf hin, daß die Daten zwar der Eigenschaft MAR, jedoch nicht der Eigenschaft OAR genügen, da der Grund für das Fehlen weniger auf die an sich unbekannten Realisierungen der fehlenden Werte zurückgeführt werden kann, sondern eher beim Objekt selbst zu suchen ist. Zum anderen können die Daten aber auch bezüglich eines einzelnen Merkmals oder weniger Merkmale größtenteils fehlen. Da in diesem Fall der Grund für das Fehlen vermutlich eher vom Merkmal selbst und weniger von den Objekten ausgeht, deutet dieses MD-Muster darauf hin, daß die Eigenschaft MAR nicht erfüllt ist und die Daten somit systematisch fehlen.

2.5 Konsequenzen für die Analyse der fehlenden Daten

Im folgenden sollen die aus den Ausführungen dieses Kapitels resultierenden Konsequenzen für die Analyse der fehlenden Daten diskutiert werden. Da lediglich die Auswirkungen der zu den fehlenden Werten führenden Ursachen auf das vorliegende Datenmaterial von Bedeutung und somit zu berücksichtigen sind, stellt sich die Frage nach den Abhängigkeitsbeziehungen der fehlenden Werte innerhalb einer Daten- bzw. Distanzmatrix. Dabei können die folgenden drei Fälle unterschieden werden:

- Abhängigkeit der fehlenden Daten von den an sich unbekannten Realisierungen dieser Werte
- Abhängigkeit der fehlenden Daten vom Fehlen der Daten bei anderen Merkmalen bzw. Objekten
- Abhängigkeit der fehlenden Daten von den vorhandenen oder fehlenden Werten bei anderen Merkmalen oder Objekten

Bei Vorliegen einer unvollständigen Datenmatrix können alle drei dargestellten Abhängigkeitsbeziehungen einzeln oder im Verbund auftreten. Im Fall einer unvollständigen Distanzmatrix sind nur die ersten beiden Abhängigkeitsbeziehungen zu berücksichtigen, da eine Abhängigkeit zwischen dem Fehlen einer paarweisen Distanz und dem Wert einer anderen paarweisen Distanz nicht vorstellbar ist. Im Hinblick auf die im nächsten Kapitel folgende Strukturanalyse einer unvollständigen Daten- bzw. Distanzmatrix, bei der unter anderem die möglichen Abhängigkeitsbeziehungen der fehlenden Werte aufgedeckt werden sollen, ist an dieser Stelle festzuhalten, daß eine Untersuchung der Abhängigkeit der fehlenden Daten von den fehlenden Werten bei anderen Merkmalen oder Objekten nicht möglich ist, da diese Werte unbekannt sind. Für den Fall einer zu analysierenden Abhängigkeit der fehlenden Daten von den ebenfalls unbekannten Realisierungen dieser Werte gilt diese Einschränkung nicht unbedingt, da beispielsweise die Verteilung der Grundgesamtheit als zusätzliche, externe Information für eine derartige Untersuchung herangezogen werden kann.

Im folgenden soll der für eine Analyse der fehlenden Daten bedeutende Zusammenhang zwischen den untersuchbaren Abhängigkeitsbeziehungen und den im Rahmen der unsystematischen Ausfallmechanismen vorgestellten Eigenschaften MAR und OAR für die Daten aufgezeigt werden. Liegt ausgehend von einer unvollständigen Daten- oder Distanzmatrix eine Abhängigkeit der fehlenden Daten von den an sich unbekannten Realisierungen dieser Werte vor, dann ist die gemäß (2.5) definierte Eigenschaft MAR für die Daten nicht erfüllt. Mit Ausnahme des vernachlässigbaren Falles, daß sämtliche Werte fehlen, genügen die Daten jedoch auch nicht der Bedingung OAR nach (2.7), da eine Abhängigkeit der fehlenden Daten von deren unbekannten Realisierungen den entsprechenden Zusammenhang zwischen dem Vorhandensein der Daten und deren beobachteten Werten impliziert. Im Fall einer Abhängigkeit der fehlenden Daten vom Fehlen der Daten bei anderen Merkmalen oder Objekten sind analog die Eigenschaften MAR und OAR nicht erfüllt, da die Ausfallwahrscheinlichkeit in Beziehung zu den fehlenden und damit implizit auch den vorhandenen Daten steht. Bei einer Abhängigkeit der fehlenden Daten von den vorhandenen Ausprägungen bei anderen Merkmalen oder Objekten genügen die Daten schließlich der Bedingung MAR, während die Eigenschaft OAR nicht erfüllt ist. In der Tabelle 2.1 sind noch einmal die für eine Analyse der fehlenden Daten relevanten Abhängigkeitsbeziehungen innerhalb einer Daten- bzw. Distanzmatrix und die dadurch jeweils resultierenden Eigenschaften der Daten dargestellt.

Abschließend sei noch darauf hingewiesen, daß in der Datenmatrix des Anhangs A sowie in der Distanzmatrix des Anhangs B, die zur Erläuterung der in den nachfolgenden Kapiteln 3 und 4 vorgestellten Methoden und Ansätze im Rahmen von Beispielen dienen, die in diesem Abschnitt diskutierten, für die Analyse der fehlenden Daten relevanten

Abhängigkeitsbeziehungen berücksichtigt wurden. Weitere Einzelheiten und Erläuterungen können den Anhängen A und B entnommen werden.

Abhängigkeitsbeziehung der MD	**Ausgangspunkt**	**Eigenschaft der Daten**
Abhängigkeit der MD von den an sich unbekannten Realisierungen dieser Werte	Daten- oder Distanzmatrix	Daten sind nicht MAR und nicht OAR
Abhängigkeit der MD vom Fehlen der Daten bei anderen Merkmalen bzw. Objekten	Daten- oder Distanzmatrix	Daten sind nicht MAR und nicht OAR
Abhängigkeit der MD von den vorhandenen Ausprägungen bei anderen Merkmalen oder Objekten	Datenmatrix	Daten sind MAR und nicht OAR

Tabelle 2.1: Überblick der für die Analyse der MD relevanten Abhängigkeitsbeziehungen

3 Strukturanalyse unvollständiger Daten- und Distanzmatrizen

Aufgrund der zum Teil nicht kontrollierbaren Bedingungen einer Untersuchungssituation sind die Ursachen für die fehlenden Daten meist nicht definitiv bestimmbar. Um eine unvollständige Daten- oder Distanzmatrix jedoch adäquat behandeln zu können, muß der jeweils vorliegende Ausfallmechanismus bekannt sein oder aber zumindest im Rahmen einer **Strukturanalyse** der unvollständigen Matrix näher untersucht werden.

Zuvor sollte jedoch geprüft werden, ob eine Nacherhebung der fehlenden Daten grundsätzlich möglich ist. Des weiteren ist zu klären, ob es sich tatsächlich um fehlende Daten handelt oder diese durch Nachlässigkeit oder Fehlinterpretationen bei der Codierung entstanden sind. So ist beispielsweise ein Vergleich mit den Fragebögen, der auch in Form von Stichproben durchgeführt werden kann, eine durchaus geeignete Maßnahme zur Prüfung des Datenaufbereitungsprozesses (vgl. *Lösel, Wüstendörfer, 1974, S. 344*).

Nach Abschluß aller Möglichkeiten der Datenprüfung ist nun der zu den fehlenden Daten führende Mechanismus näher zu untersuchen. Dabei sind lediglich die Auswirkungen der Ausfallursachen auf das vorliegende Datenmaterial von Bedeutung. Diese können anhand bestimmter Abhängigkeitsbeziehungen der fehlenden Werte innerhalb der vorliegenden Daten- bzw. Distanzmatrix sowie unter Berücksichtigung gewisser externer Informationen allerdings nur zum Teil untersucht werden. Damit sind die Ergebnisse einer Strukturanalyse zwar hinreichend für die Annahme eines systematischen, jedoch lediglich notwendig für die Annahme eines unsystematischen Ausfallmechanismus. Aus diesem Grund sollten im Rahmen einer Strukturanalyse der vorliegenden unvollständigen Daten- bzw. Distanzmatrix möglichst viele unterschiedliche Ansätze und Verfahren zur Anwendung kommen. Nur wenn kein Ergebnis gegen die Annahme eines unsystematischen Ausfallmechanismus spricht, ist diese Annahme auch für die weitere Analyse tragbar.

In **Abschnitt 3.1** erfolgt zunächst eine Darstellung deskriptiver Verfahren zur Analyse der Struktur unvollständiger Daten- und Distanzmatrizen. Diese Verfahren geben lediglich einen ersten Einblick in die Struktur des unvollständigen Datenmaterials und besitzen nur eine geringe Aussagefähigkeit über den zugrundeliegenden Ausfallmechanismus. Daher werden in **Abschnitt 3.2** anschließend die Möglichkeiten einer explorativen Analyse, also der Suche nach Zusammenhängen innerhalb einer unvollständigen Daten- oder Distanzmatrix, aufgezeigt, während sich der **Abschnitt 3.3** mit einer induktiven Analyse, also der Überprüfung vorher formulierter Hypothesen bezüglich der fehlenden Daten, beschäftigt. Vor allem mit den zur induktiven Analyse zählenden An-

sätzen und Methoden ist eine explizite Untersuchung des Ausfallmechanismus möglich. Der **Abschnitt 3.4** widmet sich schließlich den Konsequenzen, die sich aus den erhaltenen Ergebnissen der durchgeführten Strukturanalyse für die weitere datenanalytische Untersuchung und damit für die Behandlung der fehlenden Daten ergeben.

Die in der Literatur beschriebenen Verfahren zur Strukturanalyse im Fall fehlender Daten gehen ausnahmslos von einer unvollständigen Datenmatrix aus. Diese Ansätze können jedoch zum Teil auch bei Vorliegen einer unvollständigen, unmittelbar erhobenen Distanzmatrix in modifizierter Form angewandt werden. Bei der Darstellung der einzelnen Verfahren wird diese Tatsache entsprechend berücksichtigt.

3.1 Deskriptive Analyse

Neben der Berechnung möglichst aussagekräftiger Kennzahlen kann die Struktur des unvollständigen Datenmaterials grafisch veranschaulicht werden. Diese beiden Verfahrensvarianten, die erste Ansatzpunkte im Rahmen einer umfassenden Untersuchung des vorliegenden Ausfallmechanismus darstellen, sollen im folgenden behandelt werden.

3.1.1 Missing-Data-Maße

Den Ausgangspunkt der Betrachtung stellt im Fall einer unvollständigen Datenmatrix die Indikatormatrix V gemäß (2.3) und im Fall einer unvollständigen, unmittelbar erhobenen Distanzmatrix die Indikatormatrix W gemäß (2.4) dar. Eine Verdichtung der in diesen Matrizen enthaltenen Informationen kann durch sogenannte **Missing-Data-Maße** oder auch kurz **MD-Maße** erreicht werden. Ein MD-Maß stellt eine Kennzahl dar, die das Vorliegen, das Ausmaß oder eventuelle Konzentrationstendenzen der fehlenden Werte innerhalb der Daten- bzw. Distanzmatrix zum Ausdruck bringt. Die MD-Maße für einzelne Objekte oder Merkmale können dabei auch in sogenannten **MD-Variablen** zusammengefaßt werden. Nachfolgend wird zwischen einfachen Kennzahlen und Zusammenhangsmaßen unterschieden. Bei den Zusammenhangsmaßen werden im Gegensatz zu den einfachen Kennzahlen die Beziehungen der fehlenden Daten zwischen einzelnen Objekten bzw. Merkmalen berücksichtigt.

3.1.1.1 Einfache Kennzahlen

Im Fall einer unvollständigen Datenmatrix können zunächst die in der Tabelle 3.1 dargestellten Missing-Data-Maße berechnet werden. Diese Kennzahlen, bei denen es sich

ausschließlich um absolute Werte handelt, werden darüber hinaus als grundlegende Symbolik für die weiteren Betrachtungen in diesem Kapitel verwendet.

Missing-Data-Maß	**Bezeichnung**	
$v_{i\bullet}^{ind} = \begin{cases} 1 & \text{falls } v_{ik} = 1 \ \forall \ k \in M \\ 0 & \text{sonst} \end{cases} \quad (i = 1,\ldots,n)$	MD-Indikator für Objekt i	(3.1)
$v_{\bullet k}^{ind} = \begin{cases} 1 & \text{falls } v_{ik} = 1 \ \forall \ i \in N \\ 0 & \text{sonst} \end{cases} \quad (k = 1,\ldots,m)$	MD-Indikator für Merkmal k	(3.2)
$v_{i\bullet}^{mis} = m - \sum_{k=1}^{m} v_{ik} \quad (i = 1,\ldots,n)$	Anzahl der fehlenden Daten bei Objekt i	(3.3)
$v_{\bullet k}^{mis} = n - \sum_{i=1}^{n} v_{ik} \quad (k = 1,\ldots,m)$	Anzahl der fehlenden Daten bei Merkmal k	(3.4)
$v_{i\bullet}^{obs} = \sum_{k=1}^{m} v_{ik} = m - v_{i\bullet}^{mis} \quad (i = 1,\ldots,n)$	Anzahl der vorhandenen Daten bei Objekt i	(3.5)
$v_{\bullet k}^{obs} = \sum_{i=1}^{n} v_{ik} = n - v_{\bullet k}^{mis} \quad (k = 1,\ldots,m)$	Anzahl der vorhandenen Daten bei Merkmal k	(3.6)
$v^{mis} = \sum_{i=1}^{n} v_{i\bullet}^{mis} = \sum_{k=1}^{m} v_{\bullet k}^{mis}$	Anzahl der fehlenden Daten in der Datenmatrix	(3.7)
$v^{obs} = \sum_{i=1}^{n} v_{i\bullet}^{obs} = \sum_{k=1}^{m} v_{\bullet k}^{obs}$	Anzahl der vorhandenen Daten in der Datenmatrix	(3.8)

Tabelle 3.1: Absolute Kennzahlen für eine unvollständige Datenmatrix

Die in (3.1) definierte, binäre MD-Indikatorvariable wird von *Cohen und Cohen (1975, S. 287-289)* vorgeschlagen. Sie nimmt den Wert 1 an, falls bei einem Objekt $i \in N$ bezüglich der erhobenen Merkmale alle Werte vorliegen. Der Punkt bei $v_{i\bullet}^{ind}$ deutet die Betrachtung aller Merkmale $k \in M$ und der Index *ind* die Indikatoreigenschaft an. Analog kann gemäß (3.2) eine binäre MD-Indikatorvariable für die Merkmale bestimmt werden. In diesem Fall deutet der Punkt bei $v_{\bullet k}^{ind}$ die Betrachtung aller Objekte $i \in N$ an.

Rummel (1970, S.266) schlägt die Berechnung einer MD-Variable, die das Ausmaß an fehlenden Daten für jedes Objekt $i \in N$ angibt, gemäß (3.3) vor. Dieser Ansatz kann dahingehend modifiziert werden, daß eine MD-Variable gemäß (3.4), deren Wert die Anzahl der fehlenden Daten für jedes Merkmal $k \in M$ angibt, ermittelt wird. In Anlehnung an (3.3) und (3.4) können natürlich auch MD-Variablen, die das Ausmaß an fehlenden Daten für jedes Objekt $i \in N$ oder jedes Merkmal $k \in M$ indirekt angeben, gemäß (3.5)

und (3.6) bestimmt werden. Diese Werte entsprechen den Randauszählungen der Matrix V (vgl. *Schnell, 1986, S. 13*).

Die Anzahl der fehlenden Daten der Datenmatrix A, die mit v^{mis} bezeichnet wird, kann durch die Addition aller nach (3.3) oder (3.4) berechneten Werte gemäß (3.7) ermittelt werden. Analog ergibt sich die Anzahl der vorhandenen Daten der Datenmatrix A gemäß (3.8). Es gilt natürlich die Beziehung $v^{mis} + v^{obs} = n \cdot m$ zwischen den Werten aus (3.7) und (3.8).

Neben der absoluten Anzahl der fehlenden bzw. vorhandenen Daten für die einzelnen Objekte bzw. Merkmale sowie für die gesamte Datenmatrix können auch jeweils die zugehörigen relativen Größen berechnet werden. Die resultierenden Kennzahlen sind in der Tabelle 3.2 zusammengefaßt. Das Vorliegen eines relativen Wertes wird dabei durch das Symbol ~ angedeutet.

Missing-Data-Maß	**Bezeichnung**	
$\tilde{v}_{i\bullet}^{mis} = \frac{1}{m} \cdot v_{i\bullet}^{mis} = 1 - \frac{1}{m} \sum_{k=1}^{m} v_{ik} \quad (i = 1,...,n)$	Anteil der fehlenden Daten bei Objekt i	(3.9)
$\tilde{v}_{\bullet k}^{mis} = \frac{1}{n} \cdot v_{\bullet k}^{mis} = 1 - \frac{1}{n} \sum_{i=1}^{n} v_{ik} \quad (k = 1,...,m)$	Anteil der fehlenden Daten bei Merkmal k	(3.10)
$\tilde{v}_{i\bullet}^{obs} = \frac{1}{m} \cdot v_{i\bullet}^{obs} = \frac{1}{m} \cdot \sum_{k=1}^{m} v_{ik} \quad (i = 1,...,n)$	Anteil der vorhandenen Daten bei Objekt i	(3.11)
$\tilde{v}_{\bullet k}^{obs} = \frac{1}{n} \cdot v_{\bullet k}^{obs} = \frac{1}{n} \cdot \sum_{i=1}^{n} v_{ik} \quad (k = 1,...,m)$	Anteil der vorhandenen Daten bei Merkmal k	(3.12)
$\tilde{v}^{mis} = \frac{1}{n \cdot m} \cdot v^{mis} = \frac{1}{n \cdot m} \cdot \sum_{i=1}^{n} v_{i\bullet}^{mis} = \frac{1}{n \cdot m} \cdot \sum_{k=1}^{m} v_{\bullet k}^{mis}$	Anteil der fehlenden Daten in der Datenmatrix	(3.13)
$\tilde{v}^{obs} = \frac{1}{n \cdot m} \cdot v^{obs} = \frac{1}{n \cdot m} \cdot \sum_{i=1}^{n} v_{i\bullet}^{obs} = \frac{1}{n \cdot m} \cdot \sum_{k=1}^{m} v_{\bullet k}^{obs}$	Anteil der vorhandenen Daten in der Datenmatrix	(3.14)

Tabelle 3.2: Relative Kennzahlen für eine unvollständige Datenmatrix

Beispiel:

Für die Datenmatrix des Anhangs A erfolgt in diesem Beispiel ausschließlich eine Betrachtung des Falls zufällig fehlender Werte (Fall 1), wobei die Berechnungen analog für die im Anhang A dargestellten Fälle systematisch fehlender Daten durchgeführt werden können. Gemäß den Formeln (3.1) bis (3.6) ergeben sich die folgenden Werte für die MD-Indikatorvariablen sowie die Anzahl der fehlenden und vorhandenen Daten der einzelnen Objekte und Merkmale:

Objekt i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$v_{i\bullet}^{ind}$	1	0	1	0	1	1	1	1	0	0	0	0	1	0	0
$v_{i\bullet}^{mis}$	0	2	0	3	0	0	0	0	1	1	1	2	0	2	1
$v_{i\bullet}^{obs}$	9	7	9	6	9	9	9	9	8	8	8	7	9	7	8

Merkmal k	1	2	3	4	5	6	7	8	9
$v_{\bullet k}^{ind}$	0	0	0	0	1	1	0	1	0
$v_{\bullet k}^{mis}$	2	1	2	3	0	0	3	0	2
$v_{\bullet k}^{obs}$	13	14	13	12	15	15	12	15	13

Die Anzahl der insgesamt fehlenden bzw. insgesamt vorhandenen Daten ergibt sich gemäß (3.7) bzw. (3.8) mit $v^{mis} = 13$ bzw. $v^{obs} = 122$. Für die Anteile der fehlenden und vorhandenen Daten der einzelnen Objekte und Merkmale erhält man gemäß den Formeln (3.9) bis (3.12) die folgenden Werte:

Objekt i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$\tilde{v}_{i\bullet}^{mis}$	0.00	0.22	0.00	0.33	0.00	0.00	0.00	0.00	0.11	0.11	0.11	0.22	0.00	0.22	0.11
$\tilde{v}_{i\bullet}^{obs}$	1.00	0.78	1.00	0.67	1.00	1.00	1.00	1.00	0.89	0.89	0.89	0.78	1.00	0.78	0.89

Merkmal k	1	2	3	4	5	6	7	8	9
$\tilde{v}_{\bullet k}^{mis}$	0.13	0.07	0.13	0.20	0.00	0.00	0.20	0.00	0.13
$\tilde{v}_{\bullet k}^{obs}$	0.87	0.93	0.87	0.80	1.00	1.00	0.80	1.00	0.87

Der nach der Formel (3.13) bzw. (3.14) berechnete Anteil der insgesamt fehlenden bzw. insgesamt vorhandenen Daten beträgt 0.0963 ($=\tilde{v}^{mis}$) bzw. 0.9037 ($=\tilde{v}^{obs}$).

Die bislang für eine unvollständige Datenmatrix vorgestellten Missing-Data-Maße sollen im folgenden auf den Fall einer unvollständigen, unmittelbar erhobenen Distanzmatrix D übertragen werden. Aufgrund der speziellen Struktur einer Distanzmatrix führen die zeilenweise und die spaltenweise Betrachtung von D und damit auch von W bei der Bestimmung von MD-Maßen zu einem identischen Ergebnis. In beiden Fällen erfolgt eine objektweise Betrachtung. Des weiteren sind die Distanzen d_{ii}, die für alle $i \in N$ gleich Null sind, per Definition vorhanden und können daher in den MD-Maßen unberücksichtigt bleiben.

In den Tabellen 3.3 und 3.4 sind die absoluten und relativen Kennzahlen für eine unvollständige, unmittelbar erhobene Distanzmatrix zusammengefaßt. Diese Kennzahlen ermöglichen eine einfache Beschreibung der fehlenden Distanzen und dienen vor allem als Symbolik für die weiteren Betrachtungen in diesem Kapitel. Aufgrund der Symmetrieeigenschaft von D kommen alle nicht auf der Hauptdiagonalen liegenden Distanzen

doppelt vor. Diese Tatsache wird in den Formeln (3.18) und (3.19) jedoch durch eine entsprechende Korrektur berücksichtigt. Es gelten die Beziehungen $w^{mis} + w^{obs} = \frac{1}{2} \cdot n \cdot (n-1)$, $\tilde{w}_i^{mis} + \tilde{w}_i^{obs} = 1 \; \forall \, i \in N$ und $\tilde{w}^{mis} + \tilde{w}^{obs} = 1$.

Missing-Data-Maß	**Bezeichnung**	
$w_i^{ind} = \begin{cases} 1 & \text{falls } w_{ij} = 1 \;\; \forall \, j \in N, j \neq i \\ 0 & \text{sonst} \end{cases} \quad (i = 1,\ldots,n)$	MD-Indikator für Objekt i	(3.15)
$w_i^{mis} = (n-1) - \sum_{j \in N, j \neq i} w_{ij} \quad (i = 1,\ldots,n)$	Anzahl der fehlenden Distanzen bei Objekt i	(3.16)
$w_i^{obs} = \sum_{j \in N, j \neq i} w_{ij} = (n-1) - w_i^{mis} \quad (i = 1,\ldots,n)$	Anzahl der vorhandenen Distanzen bei Objekt i	(3.17)
$w^{mis} = \frac{1}{2} \sum_{i=1}^{n} w_i^{mis}$	Anzahl der fehlenden Distanzen in der Distanzmatrix	(3.18)
$w^{obs} = \frac{1}{2} \sum_{i=1}^{n} w_i^{obs}$	Anzahl der vorhandenen Distanzen in der Distanzmatrix	(3.19)

Tabelle 3.3: Absolute Kennzahlen für eine unvollständige Distanzmatrix

Missing-Data-Maß	**Bezeichnung**	
$\tilde{w}_i^{mis} = \frac{1}{(n-1)} \cdot w_i^{mis} = 1 - \frac{1}{(n-1)} \cdot \sum_{j \in N, j \neq i} w_{ij} \quad (i = 1,\ldots,n)$	Anteil der fehlenden Distanzen bei Objekt i	(3.20)
$\tilde{w}_i^{obs} = \frac{1}{(n-1)} \cdot w_i^{obs} = \frac{1}{(n-1)} \cdot \sum_{j \in N, j \neq i} w_{ij} \quad (i = 1,\ldots,n)$	Anteil der vorhandenen Distanzen bei Objekt i	(3.21)
$\tilde{w}^{mis} = \frac{2}{n \cdot (n-1)} \cdot w^{mis} = \frac{1}{n \cdot (n-1)} \cdot \sum_{i=1}^{n} w_i^{mis}$	Anteil der fehlenden Distanzen in der Distanzmatrix	(3.22)
$\tilde{w}^{obs} = \frac{2}{n \cdot (n-1)} \cdot w^{obs} = \frac{1}{n \cdot (n-1)} \cdot \sum_{i=1}^{n} w_i^{obs}$	Anteil der vorhandenen Distanzen in der Distanzmatrix	(3.23)

Tabelle 3.4: Relative Kennzahlen für eine unvollständige Distanzmatrix

Beispiel:

Für die Distanzmatrix des Anhangs B erfolgt in diesem Beispiel lediglich eine Betrachtung des Falls zufällig fehlender Werte (Fall 1). Für die im Anhang B dargestellten Fälle systematisch fehlender Daten können die Berechnungen jedoch analog durchgeführt werden. Gemäß den Formeln (3.15) bis (3.17) sowie

(3.20) und (3.21) ergeben sich die folgenden Werte für die MD-Indikatorvariable sowie die Anzahl und die Anteile der fehlenden und vorhandenen paarweisen Distanzen der einzelnen Objekte:

Objekt i	1	2	3	4	5	6	7	8	9	10
w_i^{ind}	0	0	0	0	0	0	0	1	0	0
w_i^{mis}	2	2	2	2	2	2	2	0	2	2
w_i^{obs}	7	7	7	7	7	7	7	9	7	7
$\tilde{w}_i^{mis}$	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0	0.22	0.22
$\tilde{w}_i^{obs}$	0.78	0.78	0.78	0.78	0.78	0.78	0.78	1	0.78	0.78

Die Anzahl der insgesamt fehlenden bzw. insgesamt vorhandenen paarweisen Distanzen ergibt sich gemäß (3.18) bzw. (3.19) mit $w^{mis} = 9$ bzw. $w^{obs} = 36$. Der nach der Formel (3.22) bzw. (3.23) berechnete Anteil der insgesamt fehlenden bzw. insgesamt vorhandenen Distanzen beträgt 0.2 ($=\tilde{w}^{mis}$) bzw. 0.8 ($=\tilde{w}^{obs}$).

3.1.1.2 Zusammenhangsmaße

Die Auswertung einer Datenmatrix erfolgt zum einen häufig unter Anwendung bivariater Verfahren und zum anderen stellen einige dieser Ergebnisse, wie beispielsweise Korrelationskoeffizienten, den Ausgangspunkt für eine Reihe multivariater Verfahren der Datenanalyse dar. Deshalb wird von *Brown (1983, S. 282-284)* ein MD-Maß vorgeschlagen, mit dem eine geeignete Beschreibung der fehlenden Daten bei paarweiser Betrachtung von Merkmalen möglich ist. Dieses Maß wird für ein Merkmal $k \in M$ in Anlehnung an *Brown* mit ξ_k bezeichnet und läßt sich gemäß der Formel

$$\xi_k = 1 - \frac{\sum_{l \in M, l \neq k} \sum_{i=1}^{n} \max\left\{0; (v_{ik} + v_{il} - 1)\right\}}{\sum_{l \in M, l \neq k} v_{\bullet l}^{obs}} \qquad (k = 1, \ldots, m) \tag{3.24}$$

berechnen. Die Werte $v_{\bullet l}^{obs}$ ($l \in M$) sind dabei gemäß (3.6) zu bestimmen. Werden also lediglich die paarweise vollständig vorliegenden Ausprägungen zwischen dem Merkmal k und jeweils einem der anderen Merkmale $l \in M$, $l \neq k$ betrachtet, dann wird mit ξ_k der dadurch entstehende Verlust an vorhandenen Daten bei allen Merkmale $l \in M$, $l \neq k$ zum Ausdruck gebracht.

Analog kann auch ein entsprechendes MD-Maß ξ_i für den Fall der Betrachtung aller paarweisen Objektkombinationen eines Objektes $i \in N$ mit jeweils einem der anderen Objekte $j \in N, j \neq i$ definiert werden. Diese Vorgehensweise ist vor dem Hintergrund einer denkbaren Berechnung paarweiser Objektdistanzen in einer datenanalytischen Un-

tersuchung zweckmäßig. Mit den nach (3.5) ermittelten Werten $v_{j\bullet}^{obs}$ ($j \in N$) ergibt sich ξ_i gemäß der Formel

$$\xi_i = 1 - \frac{\sum_{j \in N, j \neq i} \sum_{k=1}^{m} \max\left\{0; \left(v_{ik} + v_{jk} - 1\right)\right\}}{\sum_{j \in N, j \neq i} v_{j\bullet}^{obs}} \quad (i = 1,...,n). \tag{3.25}$$

Die nach (3.24) und (3.25) berechneten Maße liegen immer im Intervall $[0,1]$. Ein Wert von 0 bedeutet, daß keine vorhandenen Daten der Merkmale $l \in M$, $l \neq k$ bzw. der Objekte $j \in N$, $j \neq i$ verloren gehen. Bei einem Wert von 1 liegt hingegen kein einziges vollständiges Ausprägungspaar, das mit dem Merkmal $k \in M$ und jeweils einem der anderen Merkmale $l \in M$, $l \neq k$ bzw. mit dem Objekt $i \in N$ und jeweils einem der anderen Objekte $j \in N, j \neq i$ gebildet werden kann, vor.

Der Nachteil, daß lediglich der Verlust an vorhandenen Daten der Merkmale $l \in M$, $l \neq k$ bzw. der Objekte $j \in N, j \neq i$ zum Ausdruck gebracht wird, kann durch Verwendung von entsprechend korrigierten Maßen ξ_k^{kor} bzw. ξ_i^{kor}, die sich gemäß den Formeln

$$\xi_k^{kor} = 1 - \frac{\sum_{l \in M, l \neq k} \sum_{i=1}^{n} \max\left\{0; \left(v_{ik} + v_{il} - 1\right)\right\}}{\sum_{l \in M, l \neq k} \max\left\{v_{\bullet k}^{obs}; v_{\bullet l}^{obs}\right\}} \quad (k = 1,...,m), \tag{3.26}$$

$$\xi_i^{kor} = 1 - \frac{\sum_{j \in N, j \neq i} \sum_{k=1}^{m} \max\left\{0; \left(v_{ik} + v_{jk} - 1\right)\right\}}{\sum_{j \in N, j \neq i} \max\left\{v_{i\bullet}^{obs}; v_{j\bullet}^{obs}\right\}} \quad (i = 1,...,n) \tag{3.27}$$

berechnen lassen, umgangen werden. In diesem Fall wird auch der Verlust an vorhandenen Daten des Merkmals $k \in M$ bzw. des Objekts $i \in N$ berücksichtigt.

Die mit den Maßen ξ_k^{kor} und ξ_i^{kor} für einzelne Merkmale bzw. einzelne Objekte angegebenen Größen können auch für alle Merkmale bzw. alle Objekte jeweils in einem MD-Maß zusammengefaßt werden. Auf diese Weise kann der mittlere Verlust an vorhandenen Daten im Fall einer Betrachtung der paarweise vorliegenden Ausprägungen zwischen den einzelnen Merkmalen der Merkmalsmenge bzw. den einzelnen Objekten der Objektmenge zum Ausdruck gebracht werden. Unter Verwendung der Indizes M und N, die eine entsprechende Zusammenfassung über alle Merkmale der Merkmalsmenge M bzw. Objekte der Objektmenge N andeuten, ergeben sich diese MD-Maße gemäß

$$\xi_M^{kor} = 1 - \frac{\sum_{k,l \in M, l \neq k} \sum_{i=1}^{n} \max\left\{0; (v_{ik} + v_{il} - 1)\right\}}{\sum_{k,l \in M, l \neq k} \max\left\{v_{\bullet k}^{obs}; v_{\bullet l}^{obs}\right\}}, \tag{3.28}$$

$$\xi_N^{kor} = 1 - \frac{\sum_{i,j \in N, j \neq i} \sum_{k=1}^{m} \max\left\{0; (v_{ik} + v_{jk} - 1)\right\}}{\sum_{i,j \in N, j \neq i} \max\left\{v_{i\bullet}^{obs}; v_{j\bullet}^{obs}\right\}}. \tag{3.29}$$

Beispiel:

Für die Datenmatrix des Anhangs A wird wiederum nur der Fall zufällig fehlender Daten (Fall 1) betrachtet. Exemplarisch ergibt sich die Berechnung des MD-Maßes gemäß (3.25) für das Objekt 2 mit

$$\xi_2 = 1 - \frac{(7+7+4+7+7+7+7+6+7+6+6+7+6+6)}{(9+9+6+9+9+9+9+8+8+8+7+9+7+8)} = 1 - \frac{90}{115} = \frac{25}{115} \approx 0.22.$$

Unter Verwendung der Formeln (3.24) bis (3.27) erhält man die in der nachfolgenden Tabelle zusammengefaßten MD-Maße für alle Objekte und Merkmale:

Objekt i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ξ_i	0.00	0.22	0.00	0.34	0.00	0.00	0.00	0.00	0.11	0.11	0.11	0.21	0.00	0.21	0.11
ξ_i^{kor}	0.10	0.22	0.10	0.34	0.10	0.10	0.10	0.10	0.15	0.15	0.15	0.22	0.10	0.22	0.14

Merkmal k	1	2	3	4	5	6	7	8	9
ξ_k	0.14	0.06	0.13	0.19	0.00	0.00	0.20	0.00	0.13
ξ_k^{kor}	0.15	0.11	0.14	0.19	0.11	0.11	0.20	0.11	0.14

Bei den Objekten 1, 3, 5, 6, 7, 8, und 13 sowie den Merkmalen 5, 6 und 8 beträgt das unkorrigierte MD-Maß jeweils 0, da die entsprechenden Objekt- bzw. Merkmalsvektoren keine fehlenden Daten aufweisen. Im Fall einer paarweisen Betrachtung können damit die vorhandenen Ausprägungen der anderen Objekte bzw. Merkmale vollständig genutzt werden. Jedoch gehen aufgrund der fehlenden Daten der anderen Objekte bzw. Merkmale zum Teil Ausprägungen der vollständigen Objekte bzw. Merkmale verloren. Diese Tatsache wird in den korrigierten MD-Maßen entsprechend berücksichtigt.

Für die MD-Maße gemäß (3.28) und (3.29) ergeben sich im einzelnen die folgenden Werte: $\xi_M^{kor} = 0.14$ und $\xi_N^{kor} = 0.15$. Für die Datenmatrix muß damit bei Verwendung der paarweise verfügbaren Merkmalsausprägungen insgesamt ein Verlust von ca. 14 Prozent der vorhandenen Daten hingenommen werden, während bei der Verwendung der paarweise verfügbaren Objektausprägungen ein Verlust von ca. 15 Prozent der vorhandenen Daten zu verzeichnen ist.

Die Beschreibung von Konzentrationstendenzen der fehlenden Werte innerhalb einer Datenmatrix kann mittels weiterer MD-Maße erfolgen. Der Grundgedanke für diese Maße ergibt sich aus der Überlegung, daß die Objektvektoren v^i ($i \in N$) bzw. die Merkmalsvektoren v_k ($k \in M$) der Indikatormatrix V jeweils einzelne **MD-Muster** oder genauer **MD-Objektmuster** bzw. **MD-Merkmalsmuster** darstellen. Im Gegensatz zu den in Abschnitt 2.4 vorgestellten, theoretisch denkbaren Mustern fehlender Daten handelt es sich in diesem Zusammenhang um die tatsächlich vorliegenden MD-Muster der Objekte bzw. Merkmale in einer Datenmatrix. Die Konzentrationstendenzen der fehlenden Ausprägungen können nun dahingehend beschrieben werden, ob und in welchem Ausmaß einzelne MD-Muster wiederholt auftreten. In diesem Zusammenhang sind jedoch die MD-Objekt- bzw. MD-Merkmalsmuster der vollständig vorliegenden Objekte bzw. Merkmale nicht von Bedeutung, so daß im folgenden lediglich die MD-Muster der Objekte bzw. Merkmale mit fehlenden Werten betrachtet werden. Die Maximalzahl unterschiedlicher MD-Objektmuster ergibt sich dann aus dem Minimum der Anzahl der Objekte mit fehlenden Daten und dem Wert $(2^m - 1)$. Analog erhält man die Maximalzahl unterschiedlicher MD-Merkmalsmuster aus dem Minimum der Anzahl der Merkmale mit fehlenden Daten und dem Wert $(2^n - 1)$. Die Anzahl der in einer Datenmatrix tatsächlich vorliegenden unterschiedlichen MD-Objekt- bzw. MD-Merkmalsmuster kann nun jeweils im Verhältnis zur Maximalanzahl betrachtet werden, so daß sich die folgenden Maße π_N bzw. π_M ergeben, wobei der Index N bzw. M eine objekt- bzw. eine merkmalsbezogene Betrachtung andeutet:

$$\pi_N = \frac{\sum_{i=1}^{n-1}\left(1-v_{i\bullet}^{ind}\right)\cdot\left[1-\min\left\{\sum_{j=i+1}^{n}\left(1-v_{j\bullet}^{ind}\right)\cdot\left(1-\min\left\{\sum_{k=1}^{m}\left|v_{ik}-v_{jk}\right|;1\right\}\right);1\right\}\right]}{\min\left\{\left(n-\sum_{i=1}^{n}v_{i\bullet}^{ind}\right);\left(2^m-1\right)\right\}-1}, \tag{3.30}$$

$$\pi_M = \frac{\sum_{k=1}^{m-1}\left(1-v_{\bullet k}^{ind}\right)\cdot\left[1-\min\left\{\sum_{l=k+1}^{m}\left(1-v_{\bullet l}^{ind}\right)\cdot\left(1-\min\left\{\sum_{i=1}^{n}\left|v_{ik}-v_{il}\right|;1\right\}\right);1\right\}\right]}{\min\left\{\left(m-\sum_{k=1}^{m}v_{\bullet k}^{ind}\right);\left(2^n-1\right)\right\}-1}. \tag{3.31}$$

Durch den Ausdruck im Zähler, der die Anzahl der tatsächlich vorliegenden unterschiedlichen MD-Objekt- bzw. MD-Merkmalsmuster minus der Zahl 1 angibt, und die Subtraktion der Zahl 1 im Nenner von (3.30) bzw. (3.31) wird garantiert, daß die berechneten Maße das Intervall $[0,1]$ vollständig ausschöpfen. Ein Wert von 1 bedeutet dabei, daß alle betrachteten MD-Objekt- bzw. MD-Merkmalsmuster unterschiedlich sind, während ein Wert von 0 auf das Vorliegen identischer MD-Objekt- bzw. MD-Merkmalsmuster hinweist. Je mehr unterschiedliche MD-Muster vorhanden sind, d.h. je grö-

ßer π_N und π_M sind, desto eher wird man vermuten, daß die Daten zufällig fehlen. Kleine Werte für π_N und π_M deuten hingegen auf ein systematisches Fehlen der Daten hin. Damit geben diese beiden Maße einen ersten Anhaltspunkt für das Formulieren von Hypothesen bezüglich des Vorliegens systematischer bzw. unsystematischer Ausfallmechanismen, die dann im Rahmen einer induktiven Analyse überprüft werden können.

Beispiel:

Für die Datenmatrix des Anhangs A soll zunächst der dort beschriebene Fall 1, also der Fall zufällig fehlender Daten, betrachtet werden. Gemäß der Formel (3.30) ergibt sich exemplarisch die folgende Berechnung, wobei die Tabelle zur Bestimmung des Ausdrucks im Zähler dient:

i	$\left(1-v_{i\bullet}^{ind}\right)$	$\left[1-\min\left\{\sum_{j=i+1}^{n}\left(1-v_{j\bullet}^{ind}\right)\cdot\left(1-\min\left\{\sum_{k=1}^{m}\lvert v_{ik}-v_{jk}\rvert;1\right\}\right);1\right\}\right]$	j	$\left(1-v_{j\bullet}^{ind}\right)$	$\left(1-\min\left\{\sum_{k=1}^{m}\lvert v_{ik}-v_{jk}\rvert;1\right\}\right)$
1	0				
2	1	1 – min {0;1} = 1	3	0	
			4	1	1 – min {5;1} = 0
			5	0	
			6	0	
			7	0	
			8	0	
			9	1	1 – min {3;1} = 0
			10	1	1 – min {1;1} = 0
			11	1	1 – min {3;1} = 0
			12	1	1 – min {2;1} = 0
			13	0	
			14	1	1 – min {2;1} = 0
			15	1	1 – min {3;1} = 0
⋮	⋮	⋮	⋮	⋮	⋮

$$\pi_N = \frac{(0+1+0+1+0+0+0+0+1+1+1+0+0+1)}{\min\{8;511\}-1} = \frac{6}{7}.$$

Analog erhält man für π_M nach (3.31) den Wert $\pi_M = \frac{5}{5} = 1$. Dies bedeutet, daß die MD-Muster der Merkmale mit fehlenden Daten alle unterschiedlich sind und lediglich zwei identische MD-Muster der Objekte mit fehlenden Daten vorliegen. Diese Ergebnisse sprechen für ein zufälliges Fehlen der Daten. Im Vergleich dazu führt der im Anhang A beschriebene Fall 3, der eine Möglichkeit systematisch fehlender Daten beinhaltet, zu den folgenden Ergebnissen: $\pi_N = \frac{0}{4} = 0$ und $\pi_M = \frac{0}{1} = 0$.

3.1.2 Grafische Verfahren

Eine grundlegende Möglichkeit der grafischen Strukturanalyse einer unvollständigen Daten- oder Distanzmatrix ist durch die Darstellung der entsprechenden Indikatormatrix gegeben. Dabei können in Anlehnung und Ergänzung zu *Little und Smith (1987,*

S. 59-60) sowie *Schnell (1986, S. 13-14)* unter anderem die folgenden fünf Varianten unterschieden werden:

- Darstellung der **kompletten** Matrix V bzw. W
- Darstellung der **kompletten, objektweise sortierten** Matrix V bzw. W
- Darstellung der **kompletten, merkmalsweise sortierten** Matrix V^T
- Darstellung der **reduzierten, objektweise sortierten** Matrix V bzw. W
- Darstellung der **reduzierten, merkmalsweise sortierten** Matrix V^T

Bei einer grafischen Darstellung der **kompletten** Matrix V bzw. W können sowohl das Ausmaß wie auch die Konzentrationstendenzen der fehlenden Daten untersucht werden. Sind die fehlenden Werte jeweils regellos über die gesamte Matrix verteilt, dann wird man einen unsystematischen Ausfallmechanismus vermuten. Im Fall der Indikatormatrix W ist zusätzlich zu beachten, daß diese Matrix symmetrisch ist und in der Hauptdiagonalen grundsätzlich keine fehlenden Werte besitzt.

Beispiel:

Für die Datenmatrix des Anhangs A ergibt sich die in der Abbildung 3.1 gezeigte grafische Darstellung der kompletten Indikatormatrix V für den Fall zufällig fehlender Daten (Fall 1).

Objekt	i	Merkmal k								
		1	2	3	4	5	6	7	8	9
BMDP	1									
CRUNCH	2	o			o					
CSS	3									
MICROSTAT II	4		o	o						o
MINITAB	5									
NCSS	6									
P-STAT	7									
RS/1	8									
SAS	9			o						
SPSS	10	o								
STATA	11									o
STATGRAPHICS	12				o			o		
STATISTIX	13									
STATPAC GOLD	14				o			o		
SYSTAT	15							o		

Abbildung 3.1: Darstellung der kompletten Matrix V (Fall 1)

Die fehlenden Ausprägungen werden dabei mit dem Symbol o angedeutet, während die vorhandenen Ausprägungen in Form eines Leerzeichens dargestellt sind. Dadurch sind Ausmaß und eventuelle Konzentrationstendenzen der fehlenden Daten optisch gut erkennbar. Dabei kann festgestellt werden, daß die fehlenden Daten nicht in bestimmten Bereichen konzentriert und weitgehend regellos über die gesamte Matrix gestreut sind. Dieses Ergebnis spricht damit für das Vorliegen unsystematisch fehlender Daten.

Im Vergleich dazu erhält man für den im Anhang A beschriebenen Fall 3, der eine Form systematisch fehlender Daten zum Ausdruck bringt, die in der Abbildung 3.2 gezeigte grafische Darstellung der zugehörigen, kompletten Matrix *V*.

Objekt	i	Merkmal k								
		1	2	3	4	5	6	7	8	9
BMDP	1								o	o
CRUNCH	2								o	o
CSS	3								o	o
MICROSTAT II	4								o	o
MINITAB	5								o	o
NCSS	6									
P-STAT	7									
RS/1	8									
SAS	9									
SPSS	10									
STATA	11									
STATGRAPHICS	12									
STATISTIX	13									
STATPAC GOLD	14									
SYSTAT	15									

Abbildung 3.2: Darstellung der kompletten Matrix V (Fall 3)

Das in diesem Fall vorliegende systematische Fehlen der Daten wird anhand des konzentrierten Auftretens der fehlenden Werte in der rechten oberen Ecke ersichtlich.

Durch ein **objektweises Sortieren** der **kompletten** Matrix *V* bzw. *W* sowie durch ein **merkmalsweises Sortieren** der **kompletten** Matrix V^T kann eine übersichtliche Darstellung der MD-Objektmuster für alle Objekte $i \in N$ sowie der MD-Merkmalsmuster für alle Merkmale $k \in M$ erreicht werden. Zur Bestimmung einer geeigneten Reihenfolge der Zeilen der Matrix *V* schlägt *Schnell (1986, S. 14)* vor, die Objekte bzw. deren MD-Muster auf Basis der Ausprägungen einer sogenannten **Patternvariable**, die sich gemäß der Formel

$$PV_{i\bullet} = 2^m - 1 - \sum_{k=1}^{m} v_{ik} \cdot 2^{k-1} \qquad (i = 1,\ldots,n) \tag{3.32}$$

ergeben, zu sortieren.[1] Der Punkt bei $PV_{i\bullet}$ deutet eine Betrachtung über alle Merkmale $k \in M$ an. Anhand der Rangordnung der gemäß (3.32) für jedes Objekt berechneten Werte ergibt sich dann eine Reihenfolge der MD-Objektmuster, bei der das Ausmaß und die Position der fehlenden Daten berücksichtigt werden. Analog können für den Fall einer merkmalsweisen Sortierung der Matrix V^T sowie einer objektweisen Sortierung der Ma-

[1] Die hier dargestellte Formel für die Patternvariable weicht von der bei *Schnell (1986, S. 14)* angegebenen Formel hinsichtlich der Sortierrichtung ab.

trix W die Patternvariablen $PV_{\bullet k}$ und PV_i bestimmt werden. Die Ausprägungen dieser beiden Variablen ergeben sich gemäß

$$PV_{\bullet k} = 2^n - 1 - \sum_{i=1}^{n} v_{ik} \cdot 2^{i-1} \qquad (k = 1,\ldots,m), \tag{3.33}$$

$$PV_i = 2^n - 1 - \sum_{j=1}^{n} w_{ij} \cdot 2^{j-1} \qquad (i = 1,\ldots,n). \tag{3.34}$$

Bei der merkmalsweisen Sortierung wird die zu V transponierte Matrix V^T verwendet, um damit eine insgesamt einheitliche, zeilenweise Darstellung der MD-Muster zu erhalten.

Beispiel:

Für die Datenmatrix des Anhangs A ergeben sich für den Fall unsystematisch fehlender Daten (Fall 1) die grafischen Darstellungen der kompletten, objektweise sortierten sowie der kompletten, transponierten und merkmalsweise sortierten Indikatormatrix V der Abbildungen 3.3 und 3.4. Dabei sind die fehlenden Ausprägungen wiederum durch das Symbol o und die vorhanden Ausprägungen durch ein Leerzeichen angedeutet. Zusätzlich sind jeweils die gemäß (3.32) und (3.33) bestimmten Werte der Patternvariablen angegeben.

Objekt	i	$PV_{i\bullet}$	Merkmal k								
			1	2	3	4	5	6	7	8	9
BMDP	1	0									
CSS	3	0									
MINITAB	5	0									
NCSS	6	0									
P-STAT	7	0									
RS/1	8	0									
STATISTIX	13	0									
SPSS	10	1	o								
SAS	9	4			o						
CRUNCH	2	9	o			o					
SYSTAT	15	64							o		
STATGRAPHICS	12	72				o			o		
STATPAC GOLD	14	72				o			o		
STATA	11	256									o
MICROSTAT·II	4	262		o	o						o

Abbildung 3.3: Darstellung der kompletten, objektweise sortierten Matrix V (Fall 1)

Bei einer näheren Betrachtung der Abbildung 3.3 wird ersichtlich, daß zum einen eine Reihe von Objekten keine fehlenden Ausprägungen besitzen und zum anderen die MD-Muster der Objekte mit fehlenden Werten zum größten Teil sehr unterschiedlich sind. Lediglich die MD-Muster der Objekte STATGRAPHICS und STATPAC GOLD sind identisch. Diese Ergebnisse sprechen somit für einen unsystematischen Ausfallmechanismus bei objektweiser Betrachtung der Datenmatrix.

Merkmal	k	PV_{*k}	Objekt i														
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Testverfahren	5	0															
Multivariate Verfahren	6	0															
Businessgrafiken	8	0															
Benutzeroberfläche	2	8				o											
Programmierbarkeit	3	264				o					o						
Preisniveau	1	514		o								o					
Statistische Grafiken	9	1032				o							o				
Deskriptive Statistik	4	10242		o										o		o	
Spezialgebiete	7	26624												o		o	o

Abbildung 3.4: Darstellung der kompletten, merkmalsweise sortierten Matrix V^T (Fall 1)

Eine Betrachtung der Abbildung 3.4 führt zu der Vermutung, daß auch merkmalsweise ein unsystematisches Fehlen der Daten vorliegt, da die MD-Merkmalsmuster optisch weitgehend unterschiedlich sind. Auf eine Darstellung der kompletten, objekt- bzw. merkmalsweise sortierten Indikatormatrix für die im Anhang A beschriebenen Fälle systematisch fehlender Daten wird aufgrund der im Prinzip identischen Vorgehensweise an dieser Stelle verzichtet.

Bei der Darstellung der **reduzierten, objektweise sortierten** Matrix V bzw. W sowie der **reduzierten, merkmalsweise sortierten** Matrix V^T werden jeweils nach dem Sortieren mittels der entsprechenden Patternvariablen nur noch die jeweils unterschiedlichen MD-Objekt- bzw. MD-Merkmalsmuster verwendet. Zusätzlich erfolgt eine weitere Reduktion der jeweiligen Matrix auf die Spalten mit mindestens einem fehlenden Wert. Damit wird auch im Fall einer großen Anzahl von Objekten bzw. Merkmalen die Übersichtlichkeit gewährleistet. Zur besseren Interpretation der Darstellungen kann die Anzahl der Objekte bzw. Merkmale, die das entsprechende MD-Muster aufweisen, angegeben werden.

Beispiel:

Für die Datenmatrix des Anhangs A erfolgt wiederum nur eine Betrachtung des Falls zufällig fehlender Daten (Fall 1). Dabei ergeben sich die folgenden, in den Abbildung 3.5 und 3.6 gezeigten Darstellungen der reduzierten, objekt- sowie merkmalsweise sortierten Indikatormatrix:

MD-Muster	Anzahl der Objekte	Merkmal k					
		1	2	3	4	7	9
1	7						
2	1	o					
3	1			o			
4	1	o			o		
5	1					o	
6	2				o	o	
7	1						o
8	1		o	o			o

Abbildung 3.5: Darstellung der reduzierten, objektweise sortierten Matrix V (Fall 1)

MD-Muster	Anzahl der Merkmale	Objekt i							
		2	4	9	10	11	12	14	15
1	3								
2	1		o						
3	1		o	o					
4	1	o			o				
5	1		o			o			
6	1	o					o	o	
7	1						o	o	o

Abbildung 3.6: Darstellung der reduzierten, merkmalsweise sortierten Matrix V^T (Fall 1)

An dieser Stelle ist noch kritisch anzumerken, daß lediglich gewisse Formen systematisch fehlender Daten in einer grafischen Darstellung der MD-Indikatormatrix erkennbar sind. So kann unter Umständen der Fall, daß bei mehreren Merkmalen die fehlenden Daten von den Realisierungen dieser Werte abhängen, also ein systematischer Ausfallmechanismus vorliegt, zu einer weitgehend regellosen Streuung der fehlenden Werte über die gesamte Datenmatrix führen.

Eine zur grafischen Darstellung der Indikatormatrix ähnliche Methode ist die grafische Veranschaulichung der MD-Muster der Objekte bzw. Merkmale unter Verwendung von **Sterndiagrammen**. Ausgehend von einer unvollständigen Daten- bzw. Distanzmatrix werden dabei die Objektvektoren von V bzw. W, also die jeweiligen MD-Muster der Objekte, in Form von Sternen dargestellt. Bei Vorliegen einer unvollständigen Datenmatrix können auch die MD-Merkmalsmuster grafisch veranschaulicht werden. Ein vorhandener Sternzacken signalisiert dabei eine vorhandene Ausprägung bzw. Distanz, während ein nicht vorhandenen Sternzacken entsprechend eine fehlende Ausprägung bzw. Distanz andeutet. Damit ist eine visuelle, deskriptive Analyse im Hinblick auf die Ähnlichkeit der vorliegenden MD-Muster möglich. Einen unsystematischen Ausfallmechanismus wird man dann vermuten, wenn es sich mit Ausnahme der Sterne von Objekten bzw. Merkmalen ohne fehlende Daten um eine eher heterogene Menge von Sternen handelt. Eine vergleichsweise große Anzahl von ähnlichen oder sogar identischen Sternen weist hingegen auf ein systematisches Fehlen der Daten hin.

Beispiel:

Für die Datenmatrix des Anhangs A soll exemplarisch nur der Fall zufällig fehlender Daten (Fall 1) betrachtet werden. Ausgehend von den MD-Objektmustern ergeben sich für die einzelnen Objekte die in der Abbildung 3.7 gezeigten Sterne. Die Position der einzelnen Merkmale im Stern ist der Abbildung zu entnehmen.

Mit Ausnahme der Objekte ohne fehlende Daten, deren Sterne sämtliche Zacken aufweisen, und den Objekten STATGRAPHICS und STATPAC GOLD, deren Sterne zwar nicht alle Zacken besitzen, aber identisch sind, liegen weitgehend unterschiedliche Sterne und damit unterschiedliche MD-Objektmuster vor.

Eine gewisse Ähnlichkeit ist lediglich bei den Sternen der Objekte SYSTAT, STATGRAPHICS und STATPAC GOLD sowie den Objekten CRUNCH und SPSS festzustellen.

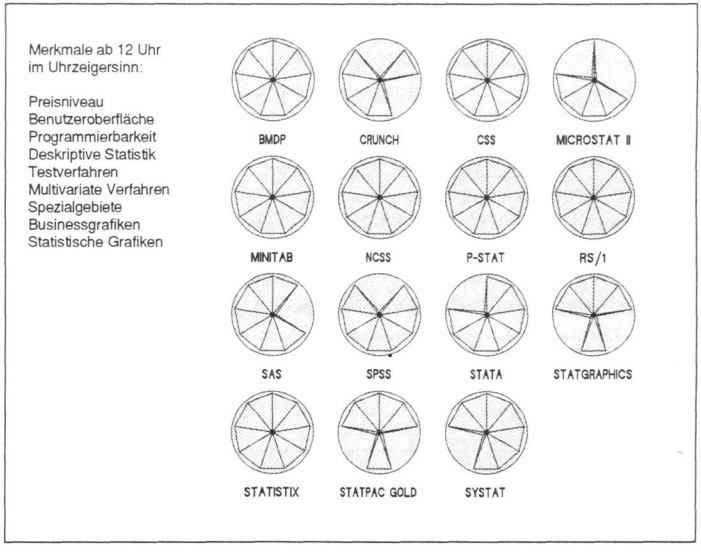

Abbildung 3.7: Sterndiagramme der MD-Objektmuster (Fall 1)

Ein weiteres grafisches Verfahren kann schließlich in Form einer Darstellung des Wertebereichs der gemäß (3.3), (3.4), (3.24) bis (3.27) und (3.16) bestimmten MD-Variablen durchgeführt werden. Dabei bietet sich gerade im Hinblick auf eine Untersuchung möglicher Ausreißer die Verwendung von **Box-Plots** an. Mit dieser grafischen Darstellung wird ein hoher Aggregationsgrad der Struktur der betrachteten Variablen erreicht. Zur Konstruktion eines Box-Plots wird zwischen dem unteren und dem oberen Quartil eine Box gezeichnet, die durch den Median unterteilt wird. Zusätzlich wird vom 10%-Quantil bis zum unteren Quartil sowie vom oberen Quartil bis zum 90%-Quantil jeweils eine Linie gezogen. Alle außerhalb dieser Bereiche liegenden Werte, die als potentielle Ausreißer zu betrachten sind, werden durch einen Kreis dargestellt (vgl. z.B. *Bausch, Opitz, 1993, S. 39*).

Beispiel:

Unter Verwendung der für die Datenmatrix des Anhangs A (Fall 1) gemäß (3.3) bzw. (3.4) bestimmten Werte, also die Anzahl der fehlenden Daten der einzelnen Objekte bzw. Merkmale, ergeben sich die in den Abbildungen 3.8 bzw. 3.9 gezeigten Box-Plots. Bei Betrachtung der MD-Anzahl der Objekte wird ersichtlich, daß ein Wert von 3 einen Ausreißer darstellt, während die Werte 0, 1 und 2 weitgehend gleichmäßig

vorliegen. Die MD-Anzahl der Merkmale ist zwar tendenziell auf Werte zwischen 0 und 2 gerichtet, jedoch stellt ein Wert von 3 keinen Ausreißer dar.

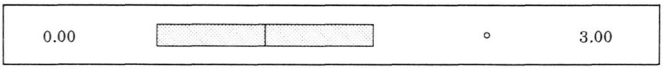

Abbildung 3.8: Box-Plot für die MD-Anzahl der Objekte der Datenmatrix (Fall 1)

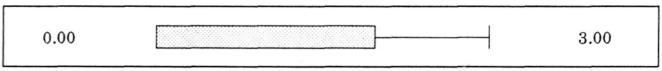

Abbildung 3.9: Box-Plot für die MD-Anzahl der Merkmale der Datenmatrix (Fall 1)

Auf eine entsprechende grafische Darstellung der gemäß (3.24) bis (3.27) für die Datenmatrix des Anhangs A berechneten Werte wird an dieser Stelle nicht eingegangen.

Für die Distanzmatrix des Anhangs B (Fall 1) ergibt sich der in der Abbildung 3.10 dargestellte Box-Plot für die gemäß (3.16) bestimmte Werte, die der Anzahl der fehlenden paarweisen Distanzen der einzelnen Objekte entsprechen.

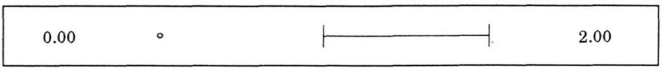

Abbildung 3.10: Box-Plot für die MD-Anzahl der Objekte der Distanzmatrix (Fall 1)

Der Box-Plot der Abbildung 3.10 verdeutlicht, daß der Wert 0 einen Ausreißer darstellt und sich die MD-Anzahl der anderen Objekte auf den Wert 2 beschränkt.

Die in diesem Abschnitt vorgestellten Methoden zur grafischen Veranschaulichung der Struktur einer unvollständigen Daten- oder Distanzmatrix besitzen nur eine begrenzte Aussagefähigkeit im Hinblick auf den zugrundeliegenden Ausfallmechanismus. Dies ist darauf zurückzuführen, daß mit den grafischen Darstellungen nicht alle im Datenmaterial grundsätzlich erkennbaren Abhängigkeitsbeziehungen der fehlenden Werte erfaßt werden können. Darüber hinaus hängt die Eignung der einzelnen Verfahren stark von der Dimension des vorliegenden Datenmaterials ab. Im Fall einer großen Anzahl von Objekten bzw. Merkmalen ist die Verwendung von Sterndiagrammen oder die Darstellung der kompletten Indikatormatrix sicherlich kein geeigneter Ansatz zur Analyse der Struktur einer unvollständigen Daten- oder Distanzmatrix. Dennoch stellen die grafischen Verfahren in vielen Fällen eine einfache Möglichkeit dar, bestimmte Abhängigkeitsbeziehungen der fehlenden Daten aufzudecken und damit erste grundlegende Erkenntnisse über den vorliegenden Ausfallmechanismus, die für die weitere Analyse herangezogen werden können, zu gewinnen.

3.2 Explorative Analyse

Im Rahmen einer explorativen Analyse soll nach Zusammenhängen innerhalb einer unvollständigen Daten- oder Distanzmatrix gesucht werden, um dadurch die gegebenenfalls vorliegenden Abhängigkeitsbeziehungen der fehlenden Werte aufzudecken. Den Ausgangspunkt der Betrachtung stellt damit, neben der Daten- bzw. Distanzmatrix, in erster Linie die jeweilige Indikatormatrix dar. Im folgenden werden die Möglichkeiten einer derartigen Analyse vorgestellt, die sich in korrelationsanalytische, faktorenanalytische, clusteranalytische sowie dependenzanalytische Ansätze einteilen lassen. In der Tabelle 3.5 sind dazu zunächst die denkbaren Analysemethoden im Hinblick auf die jeweils zu untersuchenden Abhängigkeitsbeziehungen der fehlenden Daten zusammenfassend dargestellt.

Zielsetzung	**Zielrichtung**	**Ausgangspunkt**	**Analysemethode**
Untersuchung der Abhängigkeit der MD vom Fehlen der Daten bei anderen Merkmalen bzw. Objekten	MAR, OAR	Indikatormatrix	Korrelationsanalyse, Faktorenanalyse, Clusteranalyse, Dependenzanalyse
Untersuchung der Abhängigkeit der MD von den vorhandenen Ausprägungen bei anderen Merkmalen oder Objekten	OAR	Daten- und Indikatormatrix	Korrelationsanalyse, Dependenzanalyse

Tabelle 3.5: Überblick der explorativen Analyseansätze und -methoden

Eine Untersuchung der Abhängigkeit der fehlenden Daten von den Realisierungen dieser Werte ist im Rahmen einer explorativen Analyse der Struktur unvollständiger Daten- bzw. Distanzmatrizen nicht möglich, da externe Informationen, wie beispielsweise die Kenntnis der Verteilung der Grundgesamtheit oder der Ausgangsstichprobe, nicht genutzt werden können. Damit wird deutlich, daß die Möglichkeiten einer explorativen Untersuchung im Rahmen einer umfassenden Analyse der fehlenden Daten begrenzt sind, zumal grundsätzlich die Abhängigkeitsbeziehungen der fehlenden Werte innerhalb der Daten- bzw. Distanzmatrix nur aufgezeigt und nicht statistisch überprüft werden. Dennoch sind mit den Ergebnissen einer explorativen Analyse bereits begründete Vermutungen über den zugrundeliegenden Ausfallmechanismus möglich. Die Hinweise für und gegen eine gerechtfertigte Annahme zufällig fehlender Daten geben damit grundlegende Anhaltspunkte für eine induktive Analyse der unvollständigen Daten- bzw. Distanzmatrix. Darüber hinaus können die aufgezeigten Zusammenhänge der fehlenden Werte innerhalb der Daten- bzw. Distanzmatrix für deren Behandlung im Rahmen der Datenauswertung von Bedeutung sein.

3.2.1 Korrelationsanalytische Ansätze

Zur Untersuchung der MD-Strukturen einer unvollständigen Datenmatrix schlagen *Lösel und Wüstendörfer (1974, S. 345)* die Durchführung einer Interkorrelationsanalyse der Spalten der zu A gehörenden Indikatormatrix V, also der MD-Merkmalsmuster vor. Damit wird die Abhängigkeit der fehlenden Ausprägungen vom Fehlen der Daten bei anderen Merkmalen untersucht. Die auf diese Art zu bestimmende Korrelationsmatrix, die aus den Korrelationskoeffizienten aller Paare von Merkmalen mit fehlenden Daten besteht, ergibt sich mit q als der Anzahl der Merkmale ohne fehlende Daten gemäß

$$R^{V,M} = \left(r_{kl}^{V}\right)_{m-q,m-q} \qquad \text{mit} \quad r_{kl}^{V} = \frac{\sum_{i=1}^{n}\left(v_{ik}-\overline{v}_k\right)\left(v_{il}-\overline{v}_l\right)}{\sqrt{\sum_{i=1}^{n}\left(v_{ik}-\overline{v}_k\right)^2\sum_{i=1}^{n}\left(v_{il}-\overline{v}_l\right)^2}}, \tag{3.35}$$

$$\overline{v}_k = \frac{1}{n}\sum_{i=1}^{n} v_{ik},\ \overline{v}_l = \frac{1}{n}\sum_{i=1}^{n} v_{il},\ q = \sum_{h=1}^{m} v_{\bullet h}^{ind} \quad \left(k,l \in M: v_{\bullet k}^{ind} = 0, v_{\bullet l}^{ind} = 0\right).$$

Dabei deuten der Index V den Bezug zur Indikatormatrix V und der Index M die Berechnung der Korrelationen für die Merkmalspaare an. Die Merkmale mit fehlenden Daten werden durch die Ausprägung der MD-Indikatorvariable, die jeweils nach (3.2) zu bestimmen ist, festgelegt. Analog kann auf Basis der Matrix V auch eine Korrelationsmatrix, in der die Korrelationskoeffizienten aller Paare von fehlenden Daten aufweisenden Objekten zusammengefaßt sind, berechnet werden. Diese ergibt sich gemäß

$$R^{V,N} = \left(r_{ij}^{V}\right)_{n-p,n-p} \qquad \text{mit} \quad r_{ij}^{V} = \frac{\sum_{k=1}^{m}\left(v_{ik}-\overline{v}_i\right)\left(v_{jk}-\overline{v}_j\right)}{\sqrt{\sum_{k=1}^{m}\left(v_{ik}-\overline{v}_i\right)^2\sum_{k=1}^{m}\left(v_{jk}-\overline{v}_j\right)^2}}, \tag{3.36}$$

$$\overline{v}_i = \frac{1}{m}\sum_{k=1}^{m} v_{ik},\ \overline{v}_j = \frac{1}{m}\sum_{k=1}^{m} v_{jk},\ p = \sum_{h=1}^{n} v_{h\bullet}^{ind} \quad \left(i,j \in N: v_{i\bullet}^{ind} = 0, v_{j\bullet}^{ind} = 0\right),$$

wobei p die Anzahl der Objekte ohne fehlende Daten bezeichnet und der Index N für die Berechnung der Korrelationen der Objektpaare steht. Die Ausprägungen der einzelnen Objekte im Hinblick auf die zugehörige MD-Indikatorvariable können nach der Formel (3.1) bestimmt werden.

Der von *Lösel und Wüstendörfer* vorgeschlagene Ansatz kann auch auf den Fall einer unvollständigen, unmittelbar erhobenen Distanzmatrix D übertragen werden. Den Ausgangspunkt stellt dann die Indikatormatrix W dar. Aufgrund der Tatsache, daß D und damit auch W symmetrische Matrizen sind, spielt es dabei keine Rolle, ob die Korrelationen der Zeilen oder der Spalten von W bestimmt werden. Die entsprechende Korrela-

tionsmatrix, in der die Korrelationskoeffizienten der Paare von Objekten, die zu anderen Objekten fehlende Distanzen aufweisen, zusammengefaßt sind, ergibt sich dann gemäß

$$R^W = \left(r_{ij}^W\right)_{n-o,n-o} \qquad \text{mit} \quad r_{ij}^W = \frac{\sum_{h=1}^{n}\left(w_{hi}-\overline{w}_i\right)\left(w_{hj}-\overline{w}_j\right)}{\sqrt{\sum_{h=1}^{n}\left(w_{hi}-\overline{w}_i\right)^2\sum_{h=1}^{n}\left(w_{hj}-\overline{w}_j\right)^2}}, \tag{3.37}$$

$$\overline{w}_i = \frac{1}{n}\sum_{h=1}^{n} w_{hi},\ \overline{w}_j = \frac{1}{n}\sum_{h=1}^{n} w_{hj},\ o = \sum_{h=1}^{n} w_h^{ind} \quad \left(i,j \in N: w_i^{ind} = 0, w_j^{ind} = 0\right).$$

Der Parameter o gibt die Anzahl der Objekte ohne fehlende paarweise Distanzen an und der Index W bei R^W signalisiert, daß es sich um die Korrelationsmatrix der Indikatormatrix W handelt. Die Ausprägungen der MD-Indikatorvariable ergeben sich gemäß (3.15). Durch die Summation über alle Objekte der Objektmenge in (3.37) ergibt sich zwar eine gewisse Verzerrung der berechneten Korrelationskoeffizienten, da die Werte w_{ii} und w_{ji} von den Werten w_{ij} und w_{jj} abhängen. Diese Verzerrung ist jedoch für hinreichend großes n vernachlässigbar. Eine alternative Summation in (3.37) über lediglich die Objekte $h \in N$, die ungleich den Objekten i und j sind, kann jedoch dazu führen, daß ein Korrelationskoeffizient r_{ij}^W nicht bestimmbar ist, da bei der Berechnung unter Umständen ausschließlich identische Werte w_{hi} bzw. w_{hj} verwendet werden.

Der Bravais-Pearson-Korrelationskoeffizient, der zur Bestimmung der nach (3.35) bis (3.37) berechneten Korrelationsmatrizen verwendet wird, ist grundsätzlich nur bei kardinalem Skalenniveau adäquat. Da jedoch im Fall einer Anwendung auf Ränge dieser Korrelationskoeffizient identische Werte im Vergleich zum Rangkorrelationskoeffizienten nach Spearman liefert, ergibt sich auch für das in den Indikatormatrizen vorliegende, nominal binäre Skalenniveau ein brauchbares Ergebnis (vgl. *Bausch, Opitz, 1993, S. 37*). Darüber hinaus ist alternativ auch die Anwendung des Kontingenzkoeffizienten denkbar, der im Fall nominaler Merkmale geeignet ist. Dieser Koeffizient ist beispielsweise bei *Bamberg und Baur (1993, S. 40-41)* beschrieben. Der Kontingenzkoeffizient besitzt jedoch im Vergleich zum Bravais-Pearson-Korrelationskoeffizienten den Nachteil, daß die Richtung des Zusammenhangs zwischen zwei Merkmalen nicht zum Ausdruck gebracht wird. Auf eine ausführliche Darstellung des Kontingenzkoeffizienten wird daher an dieser Stelle verzichtet.

Sind die Werte der nicht auf der Hauptdiagonalen liegenden Komponenten der nach (3.35) bis (3.37) bestimmten Korrelationsmatrizen im Bereich um Null, dann spricht dies für ein zufälliges Fehlen der Daten (vgl. *Lösel, Wüstendörfer, 1974, S.345*). Im anderen Fall können anhand hoher absoluter Werte für einzelne Korrelationskoeffizienten die vorliegenden Zusammenhänge zwischen den entsprechenden MD-Merkmals- bzw. MD-Objektmustern aufgezeigt werden. Das Vorzeichen der Koeffizienten gibt zusätzlich die

Richtung des Zusammenhangs an. Denkbare Testansätze zur Überprüfung der Hypothese, ob die gemäß (3.35) bis (3.37) bestimmten Korrelationskoeffizienten signifikant von Null abweichen, d.h. Tests auf paarweise Unabhängigkeit der Zeilen bzw. Spalten der Indikatormatrix werden im Rahmen der induktiven Analyse vorgestellt.

Als Ergänzung zu den bislang vorgestellten Ansätzen können auf Basis der Indikatormatrix die Korrelationen zwischen den Merkmalen bzw. Objekten mit fehlenden Daten und den MD-Indikatorvariablen, deren Ausprägungen gemäß (3.1), (3.2) bzw. (3.15) zu bestimmen sind, berechnet werden. Gemäß dieser Überlegung sind beispielsweise im Fall einer unvollständigen Datenmatrix bei merkmalsweiser Betrachtung die Korrelationskoeffizienten zwischen jeweils einem Spaltenvektor v_k ($k \in M$: $v_{\bullet k}^{ind} = 0$) der Indikatormatrix V und dem Vektor $\left(v_{i\bullet}^{ind}\right)_{n,1}$, der zeilenweise die nach (3.1) zu bestimmenden Ausprägungen für alle Objekte enthält, zu berechnen. Diese Koeffizienten, die im Intervall $\langle 0;1]$ liegen, können dann den zunächst nach (3.35) bis (3.37) ermittelten Korrelationsmatrizen in Form einer weiteren Zeile und Spalte hinzugefügt werden. Im Fall einer Gleichheit dieser Korrelationskoeffizienten haben alle Merkmale bzw. Objekte den gleichen Einfluß auf die insgesamt vorliegende MD-Struktur. Im Rahmen der später folgenden induktiven Analyse einer unvollständigen Daten bzw. Distanzmatrix werden entsprechende statistische Tests auf Homogenität dieser Korrelationskoeffizienten beschrieben.

Beispiel:

Für die Datenmatrix des Anhangs A ergeben sich für den Fall zufällig fehlender Daten (Fall 1) die nachfolgend dargestellten, gemäß (3.35) und (3.36) berechneten Korrelationsmatrizen, die um die jeweiligen Korrelationskoeffizienten der Objekte bzw. Merkmale auf Basis von V mit der entsprechenden MD-Indikatorvariable ergänzt sind. Diese Tatsache wird durch das Pluszeichen bei $R^{V,M+}$ bzw. $R^{V,N+}$ angedeutet. Die damit jeweils begründete zusätzliche Zeile bzw. Spalte ist durch eine gestrichelte Linie von der eigentlichen Matrix $R^{V,M}$ bzw. $R^{V,N}$ abgegrenzt.

$$R^{V,M+} = \left(\begin{array}{cccccc|c|l}
1.000 & -0.105 & -0.154 & 0.294 & -0.196 & -0.154 & 0.367 & (\text{Preisniveau}) \\
-0.105 & 1.000 & \mathbf{0.681} & -0.134 & -0.134 & \mathbf{0.681} & 0.250 & (\text{Benutzeroberfläche}) \\
-0.154 & \mathbf{0.681} & 1.000 & -0.196 & -0.196 & 0.423 & 0.367 & (\text{Programmierbarkeit}) \\
0.294 & -0.134 & -0.196 & 1.000 & 0.583 & -0.196 & 0.468 & (\text{Deskriptive Statistik}) \\
-0.196 & -0.134 & -0.196 & 0.583 & 1.000 & -0.196 & 0.468 & (\text{Spezialgebiete}) \\
-0.154 & \mathbf{0.681} & 0.423 & -0.196 & -0.196 & 1.000 & 0.367 & (\text{Statistische Grafiken}) \\
\hline
0.367 & 0.250 & 0.367 & 0.468 & 0.468 & 0.367 & 1.000 & (\text{MD - Indikatorvariable})
\end{array}\right)$$

$$R^{V,N+} = \left(\begin{array}{cccccccc|c|l}
1.000 & -0.378 & -0.189 & \mathbf{0.661} & -0.189 & 0.357 & 0.357 & -0.189 & 0.378 & (\text{CRUNCH}) \\
-0.378 & 1.000 & 0.500 & -0.250 & 0.500 & -0.378 & -0.378 & -0.250 & 0.500 & (\text{MICROSTAT II}) \\
-0.189 & 0.500 & 1.000 & -0.125 & -0.125 & -0.189 & -0.189 & -0.125 & 0.250 & (\text{SAS}) \\
\mathbf{0.661} & -0.250 & -0.125 & 1.000 & -0.125 & -0.189 & -0.189 & -0.125 & 0.250 & (\text{SPSS}) \\
-0.189 & 0.500 & -0.125 & -0.125 & 1.000 & -0.189 & -0.189 & -0.125 & 0.250 & (\text{STATA}) \\
0.357 & -0.378 & -0.189 & -0.189 & -0.189 & 1.000 & \mathbf{1.000} & \mathbf{0.661} & 0.378 & (\text{STATGRAPHICS}) \\
0.357 & -0.378 & -0.189 & -0.189 & -0.189 & \mathbf{1.000} & 1.000 & \mathbf{0.661} & 0.378 & (\text{STATPAC GOLD}) \\
-0.189 & -0.250 & -0.125 & -0.125 & -0.125 & \mathbf{0.661} & \mathbf{0.661} & 1.000 & 0.250 & (\text{SYSTAT}) \\
\hline
0.378 & 0.500 & 0.250 & 0.250 & 0.250 & 0.378 & 0.378 & 0.250 & 1.000 & (\text{MD - Indikatorvariable})
\end{array}\right)$$

Der Matrix $R^{V,M}$ liegen in der angegebenen Reihenfolge die Merkmale Preisniveau, Benutzeroberfläche, Programmierbarkeit, Deskriptive Statistik, Spezialgebiete und Statistische Grafiken zugrunde, während die Matrix $R^{V,N}$ die Korrelationskoeffizienten zwischen den Objekten CRUNCH, MICROSTAT II, SAS, SPSS, STATA, STATGRAPHICS, STATPAC GOLD und SYSTAT enthält. Die berechneten Korrelationskoeffizienten, deren absolute Werte zwar größtenteils klein sind, liegen jedoch nicht unbedingt im Bereich um Null. Aufgrund der geringen Anzahl von Objekten bzw. Merkmalen in der Datenmatrix legen diese Ergebnisse dennoch die Vermutung nahe, daß insgesamt ein unsystematisches Fehlen der Daten vorliegt. Mittels geeigneter Testverfahren, die im Rahmen einer induktiven Analyse der Struktur unvollständiger Daten- bzw. Distanzmatrizen noch vorgestellt und für dieses Beispiel ergänzt werden, kann diese Vermutung statistisch überprüft werden. Die vereinzelt vorliegenden, relativ hohen Korrelationskoeffizienten deuten auf einen Zusammenhang der entsprechenden Merkmale bzw. Objekte im Hinblick auf die fehlenden Daten hin. Dies gilt beispielsweise für die Merkmalspaare Benutzeroberfläche / Programmierbarkeit und Benutzeroberfläche / Statistische Grafiken sowie für die Objektpaare CRUNCH / SPSS, STATGRAPHICS / SYSTAT, STATPAC GOLD / SYSTAT und STATGRAPHICS / STATPAC GOLD (die entsprechenden Korrelationskoeffizienten sind in den Matrizen hervorgehoben). Anhand der Korrelationskoeffizienten der Merkmale bzw. Objekte auf Basis von V mit der entsprechenden MD-Indikatorvariable wird ersichtlich, daß alle Merkmale bzw. Objekte, die grundsätzlich fehlende Daten aufweisen, in etwa den gleichen Einfluß auf die insgesamt vorliegende MD-Struktur besitzen.

Betrachtet man den im Anhang A beschriebenen Fall 3, der die Abhängigkeit der fehlenden Daten vom Fehlen bei anderen Merkmalen berücksichtigt, dann ergeben sich die folgenden Matrizen $R^{V,M+}$ und $R^{V,N+}$:

$$R^{V,M+} = \left(\begin{array}{ccc|c} 1.000 & 1.000 & 1.000 & \text{(Businessgrafiken)} \\ 1.000 & 1.000 & 1.000 & \text{(Statistische Grafiken)} \\ \hline 1.000 & 1.000 & 1.000 & \text{(MD-Indikatorvariable)} \end{array}\right),$$

$$R^{V,N+} = \left(\begin{array}{cccccc|c} 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & \text{(BMDP)} \\ 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & \text{(CRUNCH)} \\ 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & \text{(CSS)} \\ 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & \text{(MICROSTAT II)} \\ 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & \text{(MINITAB)} \\ \hline 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & \text{(MD-Indikatorvariable)} \end{array}\right).$$

Die perfekte Abhängigkeit zwischen den Merkmalen und Objekten im Hinblick auf das Vorliegen fehlender Daten belegt den unsystematischen Ausfallmechanismus. Darüber hinaus haben alle Merkmale bzw. Objekte mit fehlenden Daten einen entscheidenden Einfluß auf die insgesamt vorliegende MD-Struktur.

Für die Distanzmatrix des Anhangs B ergibt sich zunächst für den Fall zufällig fehlender Daten (Fall 1) die folgende, gemäß (3.37) berechnete und um die Korrelationskoeffizienten der Objekte mit der entsprechenden MD-Indikatorvariable ergänzte Matrix R^{W+}:

$$R^{W+} = \left(\begin{array}{ccccccccc|c} 1.000 & 0.375 & -0.250 & -0.250 & -0.250 & 0.375 & -0.250 & -0.250 & -0.250 & 0.167 & \text{(Audi 90)} \\ 0.375 & 1.000 & 0.375 & -0.250 & -0.250 & -0.250 & -0.250 & -0.250 & -0.250 & 0.167 & \text{(BMW 3er)} \\ -0.250 & 0.375 & 1.000 & -0.250 & -0.250 & -0.250 & -0.250 & 0.375 & -0.250 & 0.167 & \text{(Ford Sierra)} \\ -0.250 & -0.250 & -0.250 & 1.000 & -0.250 & -0.250 & -0.250 & 0.375 & 0.375 & 0.167 & \text{(Mazda 626)} \\ -0.250 & -0.250 & -0.250 & -0.250 & 1.000 & 0.375 & 0.375 & -0.250 & -0.250 & 0.167 & \text{(Nissan Primera)} \\ 0.375 & -0.250 & -0.250 & -0.250 & 0.375 & 1.000 & -0.250 & -0.250 & -0.250 & 0.167 & \text{(Opel Vectra)} \\ -0.250 & -0.250 & -0.250 & -0.250 & 0.375 & -0.250 & 1.000 & -0.250 & 0.375 & 0.167 & \text{(Peugeot 405)} \\ -0.250 & -0.250 & 0.375 & 0.375 & -0.250 & -0.250 & -0.250 & 1.000 & -0.250 & 0.167 & \text{(Toyota Carina)} \\ -0.250 & -0.250 & -0.250 & 0.375 & -0.250 & -0.250 & 0.375 & -0.250 & 1.000 & 0.167 & \text{(VW Passat)} \\ \hline 0.167 & 0.167 & 0.167 & 0.167 & 0.167 & 0.167 & 0.167 & 0.167 & 0.167 & 1.000 & \text{(MD – Indikatorvariable)} \end{array}\right).$$

Der Matrix R^W liegen in der angegebenen Reihenfolge die Objekte Audi 90, BMW 3er, Ford Sierra, Mazda 626, Nissan Primera, Opel Vectra, Peugeot 405, Toyota Carina und VW Passat zugrunde. Die berechneten Korrelationskoeffizienten sprechen aufgrund der niedrigen Werte für ein zufälliges Fehlen der Distanzen. Die Korrelationskoeffizienten der Objekte auf Basis von W mit der MD-Indikatorvariable zeigen an, daß alle Objekte, die grundsätzlich fehlende Distanzen zu anderen Objekten aufweisen, den gleichen, eher geringen Einfluß auf die insgesamt vorliegende MD-Struktur besitzen. Eine Berechnung paarweiser Objektkorrelationen mit der Modifikation, daß in (3.37) eine Summation lediglich über die Objekte $h \in N$, die ungleich den Objekten i und j sind, erfolgt, führt zur folgenden Matrix

$$\tilde{R}^{W^+} = \left(\begin{array}{cccccccccc|l}
1.000 & 0.333 & -0.333 & -0.143 & -0.333 & 0.333 & -0.333 & -0.333 & -0.143 & 0.167 & \text{(Audi 90)} \\
0.333 & 1.000 & 0.333 & -0.333 & -0.333 & -0.333 & -0.143 & -0.333 & -0.143 & 0.167 & \text{(BMW 3er)} \\
-0.333 & 0.333 & 1.000 & -0.333 & -0.143 & -0.333 & -0.143 & 0.333 & -0.333 & 0.167 & \text{(Ford Sierra)} \\
-0.143 & -0.333 & -0.333 & 1.000 & -0.333 & -0.143 & -0.333 & 0.333 & 0.333 & 0.167 & \text{(Mazda 626)} \\
-0.333 & -0.333 & -0.143 & -0.333 & 1.000 & 0.333 & 0.333 & -0.143 & -0.333 & 0.167 & \text{(Nissan Primera)} \\
0.333 & -0.333 & -0.333 & -0.143 & 0.333 & 1.000 & -0.333 & -0.143 & -0.333 & 0.167 & \text{(Opel Vectra)} \\
-0.333 & -0.143 & -0.143 & -0.333 & 0.333 & -0.333 & 1.000 & -0.333 & 0.333 & 0.167 & \text{(Peugeot 405)} \\
-0.333 & -0.333 & 0.333 & 0.333 & -0.143 & -0.143 & -0.333 & 1.000 & -0.333 & 0.167 & \text{(Toyota Carina)} \\
-0.143 & -0.143 & -0.333 & 0.333 & -0.333 & -0.333 & 0.333 & -0.333 & 1.000 & 0.167 & \text{(VW Passat)} \\
\hline
0.167 & 0.167 & 0.167 & 0.167 & 0.167 & 0.167 & 0.167 & 0.167 & 0.167 & 1.000 & \text{(MD - Indikatorvariable)}
\end{array}\right)$$

Die Korrelationskoeffizienten in $\tilde{R}^{W^+}$ sind damit gegenüber den Koeffizienten in R^{W^+} um die a priori vorliegenden Abhängigkeiten der Werte w_{ii} und w_{ji} von den Werten w_{ij} und w_{jj} korrigiert. Die Interpretation verläuft analog zur Matrix R^{W^+} und führt zu denselben Ergebnissen.

Für den Fall der Abhängigkeit der fehlenden paarweisen Distanzen vom Fehlen der paarweisen Distanzen bei anderen Objekten ergeben sich für die Distanzmatrix des Anhangs B (Fall 3) die folgenden Matrizen R^{W^+} und $\tilde{R}^{W^+}$:

$$R^{W^+} = \left(\begin{array}{ccccc|l}
1.000 & 0.583 & 0.583 & 0.583 & 0.583 & 0.817 & \text{(Audi 90)} \\
0.583 & 1.000 & 0.583 & 0.583 & 0.583 & 0.817 & \text{(BMW 3er)} \\
0.583 & 0.583 & 1.000 & 0.583 & 0.583 & 0.817 & \text{(Ford Sierra)} \\
0.583 & 0.583 & 0.583 & 1.000 & 0.583 & 0.817 & \text{(Opel Vectra)} \\
0.583 & 0.583 & 0.583 & 0.583 & 1.000 & 0.817 & \text{(VW Passat)} \\
\hline
0.817 & 0.817 & 0.817 & 0.817 & 0.817 & 1.000 & \text{(MD - Indikatorvariable)}
\end{array}\right)$$

$$\tilde{R}^{W^+} = \left(\begin{array}{ccccc|l}
1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & \text{(Audi 90)} \\
1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & \text{(BMW 3er)} \\
1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & \text{(Ford Sierra)} \\
1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & \text{(Opel Vectra)} \\
1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & \text{(VW Passat)} \\
\hline
1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & \text{(MD - Indikatorvariable)}
\end{array}\right)$$

Obwohl bereits anhand der Matrix R^{W^+} das Vorliegen des systematischen Ausfallmechanismus deutlich wird, sind die perfekten Abhängigkeiten der fehlenden Distanzen vom Fehlen der Distanzen bei anderen Objekten erst in der Matrix $\tilde{R}^{W^+}$, die um die in umgekehrter Richtung, a priori vorliegenden Abhängigkeiten der Werte w_{ii} und w_{ji} von den Werten w_{ij} und w_{jj}, korrigiert ist, ersichtlich. Das Vorliegen systematisch fehlender Daten kommt auch durch die Tatsache, daß die insgesamt vorliegende MD-Struktur von allen Objekten mit fehlenden Daten entscheidend beeinflußt wird, zum Ausdruck.

Neben den bislang vorgestellten Ansätzen einer korrelationsanalytischen Untersuchung kann im Fall einer unvollständigen Datenmatrix auch eine Analyse der Abhängigkeit der fehlenden Daten von den vorhandenen Ausprägungen bei anderen Merkmalen oder

Objekten erfolgen. Dabei können im einfachsten Fall Korrelationskoeffizienten zwischen den Vektoren v_k ($k \in M$: $v_{\bullet k}^{ind} = 0$) und a_l ($l \in M$, $l \neq k$) bzw. v^i ($i \in N$: $v_{i\bullet}^{ind} = 0$) und a^j ($j \in N, j \neq i$) berechnet werden. Diese Vorgehensweise bringt jedoch einige Probleme mit sich. Zum einen ist eine objektweise Bestimmung der Korrelationskoeffizienten nur bei einheitlichem Skalenniveau aller Merkmale möglich und zum anderen ergibt sich das Problem der unter Umständen notwendigen Berücksichtigung der fehlenden Daten in den Vektoren a_l bzw. a^j. Da die ausschließliche Verwendung der vorhandenen Daten in den Vektoren a_l und a^j im Fall fehlender Werte implizit die Annahme MCAR unterstellt, ist diese Vorgehensweise im Rahmen einer Untersuchung des Vorliegens der in der Annahme MCAR enthaltenen Annahme MAR nicht korrekt. Es können also lediglich die Vektoren a_l bzw. a^j für die Berechnung der Korrelationskoeffizienten herangezogen werden, die keine fehlenden Daten enthalten. Damit ist jedoch die Effizienz dieser Analyseform im allgemeinen stark eingeschränkt. Als Korrelationskoeffizienten können abhängig vom vorliegenden Skalenniveau in den Vektoren a_l und a^j der Bravais-Pearson-Korrelationskoeffizient, der Rangkorrelationskoeffizient von Spearman sowie der Kontingenzkoeffizient zur Anwendung kommen. Da den Vektoren v_k bzw. v^i nominal binäres Skalenniveau zugrundeliegt, muß bei der Berechnung der Bravais-Pearson-Korrelation im Fall eines kardinalen Skalenniveaus in den Vektoren a_l bzw. a^j ein gewisser Fehler in Kauf genommen werden (vgl. *Bausch, Opitz, 1993, S. 37*).

Beispiel:

Für die Datenmatrix des Anhangs A ergibt sich für den dort beschriebenen Fall 4, der die Abhängigkeit der fehlenden Daten beim Merkmal Preisniveau vom Wert der Ausprägung beim Merkmal Deskriptive Statistik enthält, eine nach Bravais / Pearson berechnete Korrelation zwischen dem MD-Muster bzw. dem Indikatorvektor des Merkmals Preisniveau und dem Datenvektor des Merkmals Deskriptive Statistik von 0.906. Trotz einer gewissen Verzerrung aufgrund der unterschiedlichen Skalenniveaus in den beiden Vektoren kommt die zugrundeliegende Abhängigkeit zum Ausdruck.

Anstelle der Berechnung einzelner Korrelationskoeffizienten zwischen allen Vektoren v_k ($k \in M$: $v_{\bullet k}^{ind} = 0$) und a_l ($l \in M$, $l \neq k$) wird von *Frane (1978, S. 28)* eine **kanonische Korrelationsanalyse** zwischen der Datenmatrix A und der Indikatormatrix V vorgeschlagen. Die kanonische Korrelation zwischen allen Merkmalsvektoren aus A und allen den unvollständigen Merkmalsvektoren von A entsprechenden Spaltenvektoren aus V ergibt sich als Quadratwurzel des größten Eigenwerts des Matrizenprodukts

$$\left(S^A\right)^{-1} \cdot S^{A,V} \cdot \left(S^V\right)^{-1} \cdot \left(S^{A,V}\right)^T. \tag{3.38}$$

Dabei bezeichnen S^A die $(m \times m)$-Kovarianzmatrix bezüglich A, S^V die der Korrelationsmatrix $R^{V,M}$ aus (3.35) entsprechende $((m - q) \times (m - q))$-Kovarianzmatrix bezüglich V und $S^{A,V}$ die $(m \times (m - q))$-Kovarianzmatrix bezüglich A und V (vgl. z.B. *Hartung, Elpelt,*

1992, S.172-173). Die kanonische Korrelation ist zwar dahingehend einfach zu interpretieren, daß ein Wert nahe Null für ein unsystematisches Fehlen der Daten spricht, jedoch stellt sich bei der Berechnung der Matrizen S^A und $S^{A,V}$ das Problem fehlender Daten. Die Anwendung geeigneter Verfahren zur Berücksichtigung der fehlenden Daten bei diesen Berechnungen ist jedoch im Rahmen einer Strukturanalyse unvollständiger Datenmatrizen nicht denkbar, da eine adäquate Behandlung der fehlenden Daten erst nach erfolgter Strukturanalyse möglich ist. Damit kann insgesamt festgehalten werden, daß eine kanonische Korrelationsanalyse zur Untersuchung der Struktur einer unvollständigen Datenmatrix zumindest unter rein theoretischen Gesichtspunkten ungeeignet ist.

Alle in diesem Abschnitt vorgestellten korrelationsanalytischen Methoden zur Strukturanalyse einer unvollständigen Daten- oder Distanzmatrix setzen grundsätzlich eine hohe Anzahl an Objekten bzw. Merkmalen voraus. Im Fall einer niedrig dimensionierten Daten- bzw. Distanzmatrix können die Ergebnisse verzerrt und damit zu falschen Folgerungen bzw. Vermutungen über den zugrundeliegenden Ausfallmechanismus führen. Die denkbaren Testverfahren für die berechneten Korrelationskoeffizienten, die im Rahmen der induktiven Analyse noch vorgestellt werden, stellen eine geeignete Ergänzung einer rein explorativen, korrelationsanalytischen Untersuchung dar, da bei einer statistischen Überprüfung der zuvor formulierten Hypothesen über die Korrelationskoeffizienten die Anzahl der zugrundeliegenden Objekte bzw. Merkmale berücksichtigt wird.

3.2.2 Faktorenanalytische Ansätze

Ausgehend von den im vorherigen Abschnitt dargestellten Korrelationsmatrizen $R^{V,M}$, $R^{V,N}$ und R^W, die auf Basis der zur unvollständigen Daten- bzw. Distanzmatrix gehörenden Indikatormatrix V bzw. W bestimmt werden, kann nun eine Faktorenanalyse mittels der Hauptkomponentenmethode durchgeführt werden. Diese Vorgehensweise wird beispielsweise von *Lösel und Wüstendörfer (1974, S. 345)* vorgeschlagen. Danach dürften im Fall zufällig fehlender Daten keine substantiellen Ladungen vorliegen, da die MD-Muster der Merkmale bzw. Objekte, also die Spalten bzw. Zeilen der jeweiligen Indikatormatrix dann weitgehend unkorreliert sind. Dieser Ansatz stellt damit eine Ergänzung des korrelationsanalytischen Ansatzes zur Untersuchung der Abhängigkeit der fehlenden Daten vom Fehlen bei anderen Objekten bzw. Merkmalen dar. Die Anwendung einer Faktorenanalyse ist vor allem bei einer hohen Anzahl an Merkmalen bzw. Objekten zweckmäßig, da in diesem Fall die Korrelationsmatrizen $R^{V,M}$, $R^{V,N}$ und R^W aufgrund der ebenfalls hohen Dimension unübersichtlich werden.

Zur Erweiterung dieses Ansatzes kann ein geeignetes MD-Maß als weitere Variable in die Analyse aufgenommen wird. Diese Idee geht auf *Rummel (1970, S. 266)* zurück. Verwendet man dazu z.B. die entsprechende MD-Indikatorvariable, dann ist eine Hauptkomponentenanalyse auf Basis der im vorhergehenden Beispiel dargestellten Korrelationsmatrizen $R^{V,M+}$, $R^{V,N+}$ und R^{W+} durchzuführen. Durch niedrige Ladungen der MD-Indikatorvariable auf den Faktoren wird ersichtlich, daß das grundsätzliche Vorliegen fehlender Daten von dem Fehlen der Daten bei den einzelnen Merkmalen bzw. Objekten weitgehend unabhängig ist. Im Fall einer hohen Ladung der MD-Indikatorvariable auf einen Faktor kann entsprechend eine Abhängigkeit zwischen dem grundsätzlichen Vorliegen fehlender Daten und dem Fehlen der Daten bei den einzelnen Merkmalen bzw. Objekten, die in diesen Faktor im wesentlichen eingehen, gefolgert werden. Dabei sind jedoch die Erklärungsanteile der einzelnen Faktoren zu berücksichtigen.

Die Interpretation der Faktorladungen sowie der damit berechneten Hauptkomponenten kann durch eine orthogonale Rotation der Faktoren oder durch die Bestimmung von Korrelationen zwischen den Merkmalen bzw. Objekten der Indikatormatrizen und den Faktoren erleichtert werden (vgl. z.B. *Opitz, 1980, S. 124-125*).

Auf eine ausführliche Darstellung der Hauptkomponentenmethode, die beispielsweise bei *Weber (1974, S. 93-101)* zu finden ist, wird an dieser Stelle verzichtet. Anzumerken ist jedoch, daß die Anwendung einer Hauptkomponentenanalyse, die eigentlich von kardinalem Datenniveau ausgeht, auch bei nominal dichotomen Merkmalen, deren Ausprägungen dann metrisch behandelt werden, möglich ist (vgl. *Bausch, Opitz, 1993, S. 80*).

Beispiel:

Den Ausgangspunkt stellt die im vorherigen Beispiel für die Datenmatrix des Anhangs A im Fall zufällig fehlender Werte (Fall 1) berechnete Korrelationsmatrix $R^{V,M+}$ dar. Mittels einer Hauptkomponentenanalyse ergeben sich die folgenden Erklärungsanteile und Faktorladungen der ersten 6 Faktoren, mit denen die Ausgangsinformation vollständig erklärt wird:

Faktor	1	2	3	4	5	6
Erklärungsanteil	0.355	0.296	0.170	0.082	0.060	0.037
Preisniveau	–0.192	0.345	0.905	0.000	0.005	0.158
Benutzeroberfläche	0.879	0.147	–0.036	–0.001	0.358	0.277
Programmierbarkeit	0.821	0.126	–0.020	–0.536	–0.060	–0.135
Deskriptive Statistik	–0.373	0.795	–0.016	0.000	0.407	–0.252
Spezialgebiete	–0.327	0.692	–0.590	0.001	–0.097	0.235
Statistische Grafiken	0.820	0.126	–0.020	0.538	–0.059	–0.135
MD-Indikatorvariable	0.286	0.888	0.135	–0.001	–0.334	–0.027

Bei nur sieben in die Analyse eingehenden Merkmalen besitzt selbst der erste Faktor, der im wesentlichen durch die Merkmale Benutzeroberfläche, Programmierbarkeit und Statistische Grafiken geladen wird, lediglich einen Erklärungsanteil von 35,5 Prozent. Der daraus resultierende Zusammenhang der entsprechenden MD-Merkmalsmuster, wie auch die Abhängigkeit zwischen der MD-Indikatorvariable und den Merkmalen Deskriptive Statistik und Spezialgebiete, die durch den zweiten Faktor zum Ausdruck gebracht wird, dürfen somit nicht zu hoch bewertet werden. Bei Verwendung der ersten beiden Hauptkomponenten ergibt sich die in der Abbildung 3.11 gezeigte, grafische Darstellung der 15 Objekte auf Basis der Indikatormatrix.

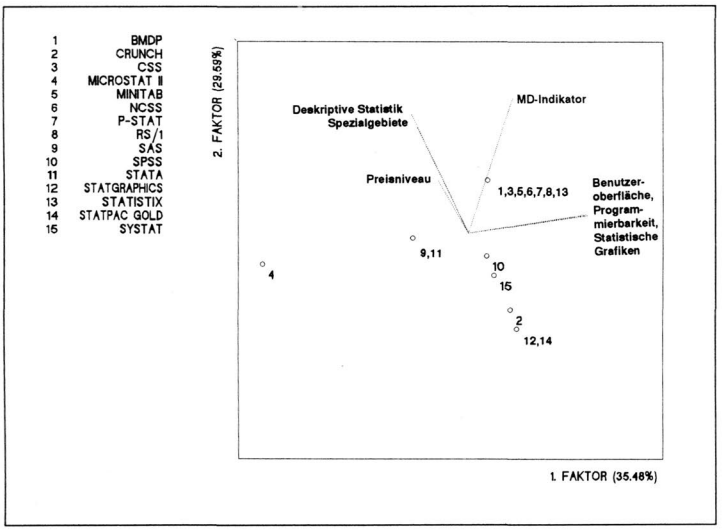

Abbildung 3.11: Zweidimensionale Darstellung der Faktorlösung

Bei einem kumulierten Erklärungsanteil von 65,1 Prozent ist die Ähnlichkeit bzw. Verschiedenheit der einzelnen MD-Objektmuster gut erkennbar. Dabei fallen alle Objekte mit vollständigen Daten, die Objekte STATGRAPHICS und STATPAC GOLD, deren MD-Muster identisch sind, sowie die Objekte STATA und SAS, deren Unterschied im MD-Muster durch eine gleichstarke Ladung der Merkmale Programmierbarkeit und Statistische Grafiken im ersten Faktor kompensiert wird, jeweils auf einen Punkt. Anhand der Merkmalsachsen wird deutlich, bezüglich welcher Merkmale die einzelnen Objekte in erster Linie fehlende Daten besitzen. So haben beispielsweise die Objekte STATGRAPHICS (12) und STATPAC GOLD (14) bezüglich der Merkmale Deskriptive Statistik und Spezialgebiet fehlende Daten, während das Objekt MICROSTAT II (4) bei den Merkmalen Benutzeroberfläche, Programmierbarkeit und Statistische Grafiken fehlende Werte aufweist. Die Tatsache, daß die MD-Indikatorvariable von keinem Merkmal entscheidend abhängt, kommt anhand der Grafik gut zum Ausdruck. Die relative Lage der Objekte entlang der Achse der MD-Indikatorvariable zeigt darüber hinaus die grundsätzliche Neigung der Objekte, fehlende Daten aufzuweisen.

Für den im Anhang A beschriebenen Fall 3 systematisch fehlender Daten erhält man ausgehend von der Korrelationsmatrix $R^{V,M+}$ mittels einer Hauptkomponentenanalyse die folgenden Erklärungsanteile und Faktorladungen aller berechneten Faktoren:

Faktor	1	2	3
Erklärungsanteil	1.000	0.000	0.000
Businessgrafiken	1.000	0.000	0.000
Statistische Grafiken	1.000	0.000	0.000
MD-Indikatorvariable	1.000	0.000	0.000

Der erste Faktor besitzt bereits einen Erklärungsanteil von 100 Prozent. Die MD-Muster der beiden Merkmale und die MD-Indikatorvariable laden diesen Faktor ohne jeglichen Informationsverlust. Die perfekte Abhängigkeit der Merkmale Businessgrafiken und Statistische Grafiken sowie die vollkommene Abhängigkeit zwischen dem grundsätzlichen Vorliegen fehlender Daten und dem Fehlen der Ausprägungen bei diesen Merkmalen kommen damit zum Ausdruck.

Auf die Darstellung einer Hauptkomponentenanalyse auf Basis der Korrelationsmatrix $R^{V,N+}$ sowie der Korrelationsmatrix R^{W+}, die sich aus der Distanzmatrix des Anhangs B ergibt, wird an dieser Stelle verzichtet.

3.2.3 Clusteranalytische Ansätze

Zur Analyse der Ähnlichkeit der MD-Muster der Merkmale bzw. Objekte und damit zur Untersuchung der Abhängigkeit der fehlenden Werte vom Fehlen der Daten bei anderen Merkmalen bzw. Objekten können Clusteranalysen durchgeführt werden (vgl. z.B. *Frane, 1978, S. 28*). Ausgehend von der zur unvollständigen Datenmatrix A bzw. unvollständigen Distanzmatrix D gehörenden Indikatormatrix V bzw. W wird also eine Klassifikation der MD-Muster der Objekte sowie, bei Betrachtung von V^T, der MD-Muster der Merkmale vorgenommen. Die Objekte bzw. Merkmale ohne fehlende Daten fallen dabei grundsätzlich in eine Klasse und können daher auch aus der Analyse ausgeschlossen werden. Falls bezüglich der Objekte bzw. Merkmale mit fehlenden Daten kein Typ eines MD-Musters überrepräsentiert ist, also bezüglich dieser Objekte bzw. Merkmale entweder keine Klassenstrukturen erkennbar sind oder eine hohe Anzahl in etwa gleich großer Klassen eine mögliche und geeignete Lösung auch unterschiedlicher Clusteranalysetechniken darstellt, deutet dies auf ein zufälliges Fehlen der Daten hin. Im Gegensatz zu den bereits vorgestellten korrelations- und faktorenanalytischen Ansätzen werden durch die Klassifikation der MD-Muster der Merkmale bzw. Objekte die Abhängigkeitsbeziehungen innerhalb der fehlenden Daten indirekt über Ähnlichkeiten bzw. Verschiedenheiten dieser MD-Muster analysiert.

Auf eine explizite Darstellung der verschiedenen Clusteranalysetechniken wird an dieser Stelle verzichtet. Einen Überblick der Klassifikationsverfahren geben beispielsweise *Backhaus et al. (1990, S. 133)*. Eine ausführliche Beschreibung der Verfahren findet sich z.B. bei *Steinhausen und Langer (1977, S. 73-157)* oder *Opitz (1980, S. 87-107)*. Die für einige Clusteranalysetechniken notwendige Bestimmung und Aggregation von Distanz- oder Ähnlichkeitsindizes ist in diesen Literaturquellen ebenfalls beschrieben (vgl. *Steinhausen, Langer, 1977, S. 51-66 ,Opitz, 1980, S. 30-64*).

Beispiel:

Die Abbildungen 3.12 und 3.13 zeigen die Dendrogramme der Complete Linkage Lösung für die MD-Muster der Objekte sowie der Single Linkage Lösung für die MD-Muster der Merkmale im Fall zufällig fehlender Werte in der Datenmatrix des Anhangs A (Fall 1).

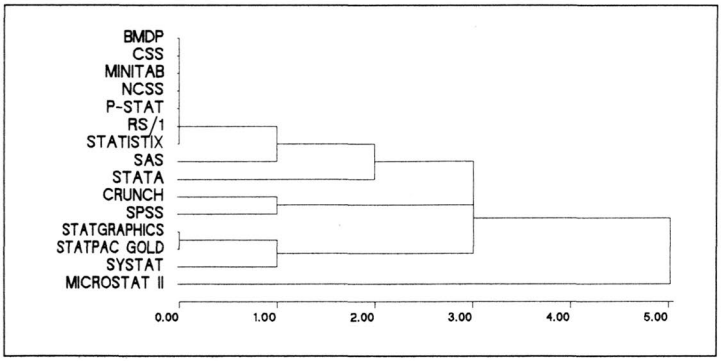

Abbildung 3.12: Complete Linkage Dendrogramm für die MD-Muster der Objekte (Fall 1)

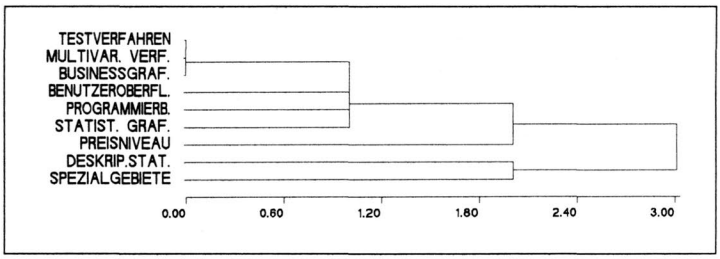

Abbildung 3.13: Single Linkage Dendrogramm für die MD-Muster der Merkmale (Fall 1)

Anhand der Abbildung 3.12 wird ersichtlich, daß sich die vorliegenden MD-Muster der Objekte mit Ausnahme der MD-Muster für die Objekte mit vollständigen Daten sowie für die Objekte STATGRAPHICS und STATPAC GOLD, deren MD-Muster identisch sind, weitgehend unterscheiden. Da sich bei jedem Übergang auf die nächstgrößere Klassenanzahl eine deutliche Verbesserung der Klassifikationsgüte er-

gibt, spricht das Complete Linkage Verfahren für ein zufälliges Fehlen der Daten. Die Durchführung der hierarchischen Klassifikationsverfahren Single Linkage und Average Linkage führt zu identischen Ergebnissen. Das in Abbildung 3.13 dargestellte Dendrogramm der Single Linkage Lösung zeigt, daß sich auch hier die Klassifikationsgüte bei jedem Übergang zur nächstgrößere Klassenzahl deutlich verbessert. Grundsätzlich können die im Rahmen der Hauptkomponentenanalyse bereits festgestellten Ähnlichkeiten einzelner MD-Merkmalsmuster zwar bestätigt werden, jedoch sind insgesamt keine, auf ein systematisches Fehlen der Daten hinweisenden Klassifikationstendenzen zu beobachten.

Für den Fall systematisch fehlender Daten in der Datenmatrix des Anhangs A (Fall 3) ergeben sich die in den Abbildungen 3.14 und 3.15 dargestellten Dendrogramme der Complete Linkage Lösung für die MD-Muster der Objekte und Merkmale.

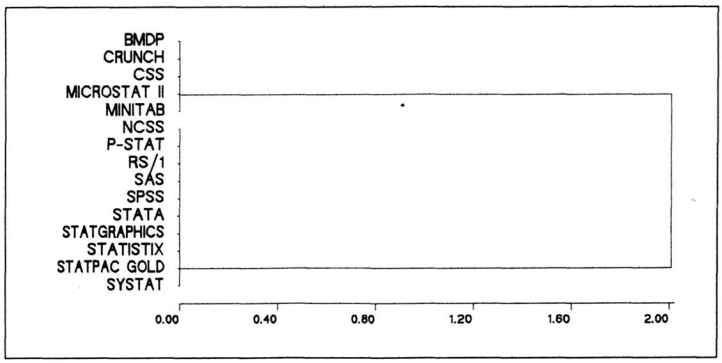

Abbildung 3.14: Complete Linkage Dendrogramm für die MD-Muster der Objekte (Fall 3)

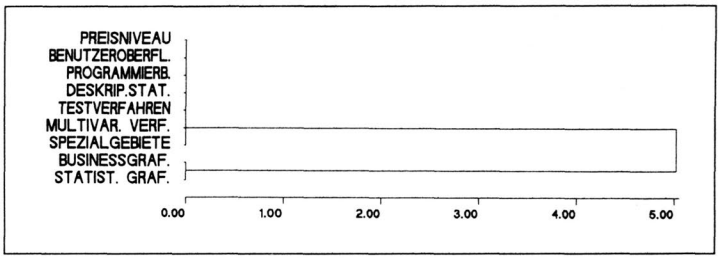

Abbildung 3.15: Complete Linkage Dendrogramm für die MD-Muster der Merkmale (Fall 3)

Die Abhängigkeit der fehlenden Ausprägungen vom Fehlen der Daten bei anderen Objekten bzw. Merkmalen kommt anhand der Tatsache zum Ausdruck, daß bei der sich ergebenden 2-Klassen-Lösung alle MD-Muster von Objekten bzw. Merkmalen mit fehlenden Daten und alle MD-Muster von Objekten bzw. Merkmalen ohne fehlende Daten jeweils eine Klasse bilden und die MD-Muster innerhalb der Klassen maximal homogen bzw. identisch sind. Die Verfahren Average Linkage und Single Linkage führen aufgrund der hier vorliegenden MD-Muster natürlich zu identischen Dendrogrammen.

Die Abbildungen 3.16 und 3.17 zeigen, ausgehend von der Distanzmatrix des Anhangs B im Fall zufällig fehlender paarweiser Distanzen (Fall 1), die Ergebnisse des Complete Linkage und Single Linkage Verfahrens für die MD-Muster der Objekte,.

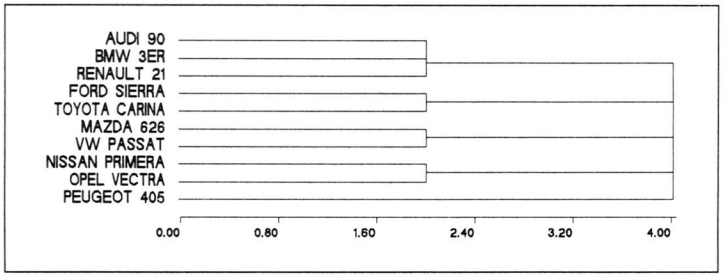

Abbildung 3.16: Complete Linkage Dendrogramm für die MD-Muster der Objekte (Fall 1)

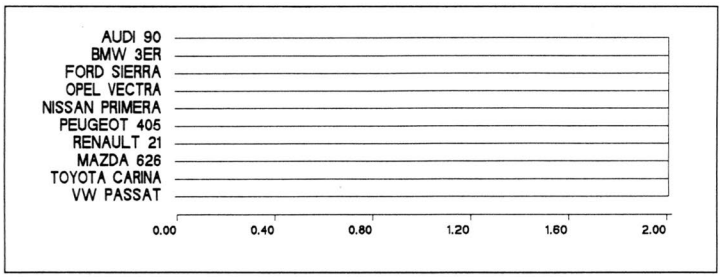

Abbildung 3.17: Single Linkage Dendrogramm für die MD-Muster der Objekte (Fall 1)

Anhand der Single Linkage Lösung sind keinerlei Klassenstrukturen zu erkennen. Gemäß dem Complete Linkage Dendrogramm ist die Bildung von fünf Klassen zweckmäßig, jedoch sind diese fünf Klassen in etwa gleich groß und die Klassenanzahl ist im Verhältnis zur Objektanzahl relativ hoch. Das zufällige Fehlen der paarweisen Distanzen wird damit zum Ausdruck gebracht.

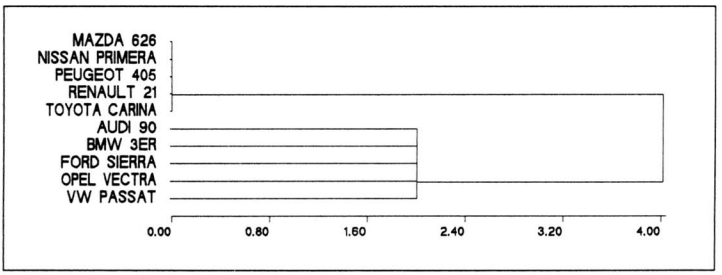

Abbildung 3.18: Single Linkage Dendrogramm für die MD-Muster der Objekte (Fall 3)

Für den Fall systematisch fehlender paarweiser Distanzen in der Distanzmatrix des Anhangs B (Fall 3) ergibt sich beispielhaft das in der Abbildung 3.18 gezeigte Dendrogramm der Single Linkage Lösung für die MD-Muster der Objekte, wobei die Verfahren Average Linkage und Complete Linkage zu identischen Darstellungen führen. Gemäß dem Dendrogramm ist die Bildung von zwei Klassen zweckmäßig. In der einen Klasse befinden sich alle Objekte ohne fehlende Werte, während die andere Klasse alle Objekte mit fehlenden paarweisen Distanzen enthält. Die anhand des Dendrogramms grundsätzlich erkennbaren Unterschiede zwischen den MD-Mustern der Objekte mit fehlenden Werten sind darauf zurückzuführen, daß sich innerhalb der Klasse von Objekten mit fehlenden Daten die MD-Muster zweier Objekte i und j im Hinblick auf die Werte w_{ii}, w_{ji}, w_{ij} und w_{jj} unterscheiden. Diese Ergebnisse sprechen damit für das Vorliegen systematisch fehlender Daten.

Da das Complete Linkage Verfahren dilatierend ist, d.h. die Tendenz zur Bildung mehrerer gleich großer Klassen besitzt, wird dieses Verfahren im allgemeinen eher auf ein zufälliges Fehlen der Daten hindeuten. Im Gegensatz zum Complete Linkage Verfahren ist das Single Linkage Verfahren kontrahierend, d.h. es weist eine Tendenz zur Kettenbildung auf. Somit wird ein systematischer Ausfallmechanismus durch das Single Linkage Verfahren im allgemeinen besser erkannt als durch das Complete Linkage Verfahren. Ein unsystematischer Ausfallmechanismus sollte also erst dann vermutet werden, wenn die Single Linkage Lösung nicht gegen das Vorliegen zufällig fehlender Daten spricht.

Abschließend wird beispielhaft eine unscharfe Klassifikation für die MD-Objektmuster der Datenmatrix des Anhangs A (Fall 1) durchgeführt. Dabei soll durch die Anwendung des bei *Bausch und Opitz (1993, S. 59-60)* beschriebenen Fuzzy-Clustering-Algorithmus eine 3-Klassen-Lösung ermittelt werden. Dieser Algorithmus setzt zwar eigentlich quantitative Daten voraus, kann aber auch im Fall eines nominal binären Skalenniveaus zur Anwendung kommen. Die Merkmalsausprägungen werden dann metrisch behandelt. In der nachfolgenden Tabelle sind die resultierenden Zugehörigkeitsgrade der Objekte zu den drei Klassen angegeben:

	Klasse 1	**Klasse 2**	**Klasse 3**
BMDP	**0.993**	0.004	0.003
CRUNCH	0.375	0.433	0.192
CSS	**0.993**	0.004	0.003
MICROSTAT II	0.054	0.035	**0.911**
MINITAB	**0.993**	0.004	0.003
NCSS	**0.993**	0.004	0.003
P-STAT	**0.993**	0.004	0.003
RS/1	**0.993**	0.004	0.003
SAS	0.456	0.175	0.369
SPSS	0.583	0.224	0.193
STATA	0.456	0.175	0.369
STATGRAPHICS	0.024	**0.964**	0.012
STATISTIX	**0.993**	0.004	0.003
STATPAC GOLD	0.024	**0.964**	0.012
SYSTAT	0.380	0.487	0.133

Anhand der Zugehörigkeitsgrade wird ersichtlich, daß in der Klasse 1 die vollständigen Objekte und in der Klasse 2 die ein identisches MD-Muster aufweisenden Objekte STATGRAPHICS und STATPAC GOLD als Kernobjekte enthalten sind. Die Klasse 3 besitzt als Kernobjekt lediglich das Objekt MICROSTAT II und damit das Objekt mit den meisten fehlenden Ausprägungen. (Die ein Kernobjekt andeutenden Zugehörigkeitsgrade sind in der Tabelle hervorgehoben.) Die restlichen Objekte weisen zwar zum Teil eine Tendenz zu einer der Klassen auf, eine eindeutige Zuordnung kann anhand der jeweiligen Zugehörigkeitsgrade jedoch nicht abgeleitet werden. Insgesamt sprechen diese Ergebnisse damit für ein zufälliges Fehlen der Daten.

3.2.4 Dependenzanalytische Ansätze

Grundsätzlich besteht das Ziel einer Dependenzanalyse in der Beschreibung und Analyse des Einflusses unabhängiger Variablen auf eine abhängige Variable. Im Rahmen einer Strukturanalyse einer unvollständigen Datenmatrix sind dabei die folgenden drei Varianten bezüglich der Wahl abhängiger und unabhängiger Variablen und damit des Analyseansatzes von Bedeutung:

- **Analyseansatz 1:** Die einzelnen merkmals- bzw. objektweisen MD-Muster werden durch die anderen merkmals- bzw. objektweisen MD-Muster erklärt, wobei die MD-Muster durch die Spalten bzw. Zeilen der Indikatormatrix V gegeben sind. Die jeweils zu analysierenden Modelle können damit folgendermaßen formuliert werden:

$$v_k = g(v_1,\ldots,v_{k-1},v_{k+1},\ldots,v_m),\ k \in M \text{ bzw. } v^i = g\left(v^1,\ldots,v^{i-1},v^{i+1},\ldots,v^n\right),\ i \in \mathrm{N}. \quad (3.39)$$

  Dabei stellt g eine beliebige reelle Funktion dar. Bei diesem Analyseansatz wird der Einfluß auf das Vorliegen fehlender Daten bei einem Merkmal bzw. Objekt durch das Vorliegen fehlender Daten bei den anderen Merkmalen bzw. Objekten untersucht.

- **Analyseansatz 2:** Die einzelnen merkmals- bzw. objektweisen MD-Muster, d.h. die Spalten bzw. Zeilen der Indikatormatrix V werden durch die Merkmals- bzw. Objektvektoren der Datenmatrix A erklärt. Damit ergeben sich die jeweils zu analysierenden Modelle gemäß

$$v_k = g(a_1,\ldots,a_{k-1},a_{k+1},\ldots,a_m),\ k \in M \text{ bzw. } v^i = g\left(a^1,\ldots,a^{i-1},a^{i+1},\ldots,a^n\right),\ i \in \mathrm{N}, \quad (3.40)$$

  wobei g wiederum eine beliebige reelle Funktion darstellt. In diesem Fall wird der Einfluß auf das Vorliegen fehlender Daten bei einem Merkmal bzw. Objekt durch die vorhandenen Daten der anderen Merkmale bzw. Objekte analysiert.

- **Analyseansatz 3:** Die einzelnen Merkmale bzw. Objekte werden durch die merkmals- bzw. objektweisen MD-Muster erklärt. Mit g als beliebiger reeller Funktion ergeben sich die jeweils zu analysierenden Modelle gemäß

$$a_k = g(v_1,\ldots,v_{k-1},v_{k+1},\ldots,v_m),\ k \in M \text{ bzw. } a^i = g(v^1,\ldots,v^{i-1},v^{i+1},\ldots,v^n),\ i \in \mathrm{N}. \quad (3.41)$$

  Bei diesem Analyseansatz wird der Einfluß auf die vorhandenen Daten eines Merkmals bzw. Objekts durch das Vorliegen fehlender Daten bei den anderen Merkmalen bzw. Objekten untersucht.

Im Fall einer objektbezogenen Strukturanalyse einer unvollständigen Datenmatrix gemäß den Analyseansätzen 2 und 3 ist noch zu ergänzen, daß ein einheitliches Skalenniveau sowie vergleichbare Ausprägungsmengen bei den Merkmalen vorliegen müssen. Bei einer Strukturanalyse einer unvollständigen Distanzmatrix ist lediglich der erste der oben vorgestellten Ansätze einer dependenzanalytischen Untersuchung, also die Erklärung der MD-Objektmuster unter Verwendung der MD-Muster der anderen Objekte, relevant. Dies ist damit zu begründen, daß eine Abhängigkeit zwischen dem Fehlen einer paarweisen Distanz und dem Wert einer anderen paarweisen Distanz prinzipiell nicht vorliegen kann.

Für die einzelnen Analyseansätze kommen je nach Skalenniveau der abhängigen und unabhängigen Variablen unterschiedliche Verfahren in Betracht. So ist beispielsweise die Diskriminanzanalyse eine grundsätzlich für den Ansatz 1, aber auch, im Fall nominal dichotomer, ordinaler und quantitativer unabhängiger Merkmale bzw. eines nominalen oder ordinalen abhängigen Merkmals, eine für die Ansätze 2 und 3 geeignete Methode. Beim Analyseansatz 3 kann im Fall eines quantitativen abhängigen Merkmals eine Varianzanalyse durchgeführt werden. Des weiteren ist bei einer auf dem Ansatz 2 basierenden Analyse auch die Anwendung einer logistischen Regression möglich. Die einzelnen Verfahren sollen an dieser Stelle nicht explizit dargestellt werden. Statt dessen wird auf die einschlägige Literatur verwiesen. So ist beispielsweise die logistische Regression bei *Toutenburg (1992, S. 233-237)*, die Diskriminanzanalyse bei *Opitz (1980, S. 146-151)* und die Varianzanalyse bei *Jobson (1991, S. 399-493)* ausführlich beschrieben. Hinweise auf die Anwendung dieser Verfahren zur Strukturanalyse einer unvollständigen Datenmatrix finden sich bei *Frane (1978, S. 28)*, *Möntmann et al. (1983, S. 95-98)* sowie *Schnell (1986, S. 16-20)*.

Für die drei dargestellten Analyseansätze ist grundsätzlich zu beachten, daß im Vergleich zur Anzahl der jeweils herangezogenen unabhängigen Variablen eine entsprechend große Anzahl an Beobachtungswerten vorliegen muß. Im Fall einer geringen Anzahl an Beobachtungswerten müssen die Modelle nach (3.39) bis (3.41) hinsichtlich der Anzahl der verwendeten unabhängigen Variablen reduziert werden.

Speziell für die Analyseansätze 2 und 3 können bei der Anwendung der dazu geeigneten Verfahren Probleme auftreten, sobald in den herangezogenen Merkmals- bzw. Objektvektoren fehlende Werte vorliegen. Da mit diesen Ansätzen die Eigenschaft OAR für die Daten überprüft werden soll, können zur Behandlung der fehlenden Werte nur die Verfahren verwendet werden, die auf der MAR-Annahme und nicht auf der restriktiveren MCAR-Annahme beruhen. Die damit in Betracht kommenden Verfahren werden in Kapitel 4 dieser Arbeit noch ausführlich vorgestellt. Auf jeden Fall unproblematisch ist die ausschließliche Verwendung vollständig vorliegender Merkmals- bzw. Objektvektoren als unabhängige bzw. abhängige Variablen.

Beispiel:

Für die Datenmatrix des Anhangs A wird zunächst für den Fall zufällig fehlender Daten (Fall 1) eine merkmalsbezogene Analyse gemäß der ersten beiden vorgestellten Ansätze durchgeführt. Aufgrund der geringen Anzahl an Objekten in der Datenmatrix werden im folgenden jeweils nur zwei unabhängige Variablen herangezogen. Betrachtet man zunächst den Analyseansatz 1, dann kann beispielsweise eine Diskriminanzanalyse mit dem MD-Muster des Merkmals Preisniveau als abhängige und den MD-Mustern der Merkmale Spezialgebiete und Statistische Grafiken als unabhängige Variablen durchgeführt werden. Die MD-Muster sind dabei jeweils durch die entsprechenden Spalten der Indikatormatrix V gegeben. Als Diskriminanzkriterium wird eine Maximierung des Quotienten aus Zwischenklassenvarianz und Innerklassenvarianzsumme gewählt. Es ergeben sich die folgenden normierten Diskriminanzkoeffizienten:

Merkmal	Spezialgebiete	Statistische Grafiken
Diskriminanzkoeffizient	0.707	0.707

Anhand der Diskriminanzkoeffizienten ist ersichtlich, daß die MD-Muster der Merkmale Spezialgebiete und Statistische Grafiken zwar einen gleichstarken Einfluß auf das MD-Muster des Merkmals Preisniveau besitzen, jedoch aufgrund des hohen Anteils an fehlerhaft klassifizierten Objekten von 53.33 Prozent insgesamt kein bedeutsamer Zusammenhang gefolgert werden kann.

Im Hinblick auf den Analyseansatz 2 soll hier beispielhaft der Einfluß der zwei vollständig erhobenen Merkmale Testverfahren und Multivariate Verfahren auf das MD-Muster des Merkmals Deskriptive Statistik untersucht werden. Im folgenden sind die Ergebnisse einer Diskriminanzanalyse auf Basis der Inversen der Innerklassen- mal Zwischenklassenkovarianzmatrix in Form der normierten Diskriminanzkoeffizienten sowie einer logistischen Regression in Form der Regressionskoeffizienten dargestellt:

Merkmal	Testverfahren	Multivariate Verfahren
Diskriminanzkoeffizient	0.738	0.675

Merkmal	Testverfahren	Multivariate Verfahren	Konstante
Regressionskoeffizient	0.064	0.054	−4.714

Für das logistische Regressionsmodell ergibt sich ein Bestimmtheitskoeffizient von 0.1641 und der Anteil der fehlerhaft klassifizierten Objekte beträgt 20 Prozent. Mit der ermittelten Diskriminanzfunktion werden 40 Prozent der Objekte fehlerhaft klassifiziert. Die Ergebnisse der beiden Verfahren zeigen im Hinblick auf die Gewichtung der beiden Merkmale eine sehr ähnliche Struktur. Insgesamt kann festgehalten werden, daß der Einfluß der Merkmale Testverfahren und Multivariate Verfahren auf das Vorliegen fehlender Daten beim Merkmal Deskriptive Statistik unbedeutend ist.

Betrachtet man für die Datenmatrix des Anhangs A den dort beschriebenen Fall 3, der die denkbare Abhängigkeit der fehlenden Daten vom Fehlen bei anderen Merkmalen berücksichtigt, dann ist der Analyseansatz 1 zur Untersuchung dieses Zusammenhangs geeignet. Nachfolgend wird eine Diskriminanzanalyse mit dem MD-Muster des Merkmals Businessgrafiken als abhängige und dem MD-Muster des Merkmale Statistische Grafiken als unabhängige Variable durchgeführt. Die MD-Muster sind jeweils durch die entsprechenden Spalten der Indikatormatrix V gegeben. Da die restlichen, vollständig vorliegenden Merkmale ohne Einfluß auf das Fehlen der Daten sind, müssen sie in der Analyse nicht berücksichtigt werden. Als Diskriminanzkriterium wird wiederum eine Maximierung des Quotienten aus Zwischenklassenvarianz und Innerklassenvarianzsumme gewählt. Für das Merkmal Statistische Grafiken ergibt sich ein normierter Diskriminanzkoeffizient von 1 bei einer fehlerfreien Identifizierung des MD-Musters des Merkmals Businessgrafiken. Damit wird der zugrundeliegende Zusammenhang zwischen diesen beiden Merkmalen deutlich.

Abschließend soll noch für den im Anhang A beschriebenen Fall 4, der die Abhängigkeit der fehlenden Daten von den vorhandenen Daten bei anderen Merkmalen berücksichtigt, der Analyseansatz 2 zur Anwendung kommen. Dabei wird der Einfluß der Merkmale Deskriptive Statistik und Testverfahren auf das MD-Muster des Merkmals Preisniveau untersucht. Man erhält die folgenden Ergebnisse einer Diskriminanzanalyse auf Basis der Inversen der Innerklassen- mal Zwischenklassenkovarianzmatrix in Form der normierten Diskriminanzkoeffizienten sowie einer logistischen Regression in Form der Regressionskoeffizienten:

Merkmal	Deskriptive Statistik	Testverfahren
Diskriminanzkoeffizient	0.985	−0.171

Merkmal	Deskriptive Statistik	Testverfahren	Konstante
Regressionskoeffizient	1.441	−0.573	−63.840

Die Diskriminanzanalyse liefert eine fehlerfrei Identifizierung des MD-Musters für das Merkmal Preisniveau. Der bedeutsame Einfluß der Ausprägungen des Merkmals Deskriptive Statistik auf das Fehlen der Ausprägungen beim Merkmal Preisniveau kommt anhand der Diskriminanzkoeffizienten zum Ausdruck. Die gleichen Ergebnisse erhält man für das logistische Regressionsmodell bei einem Bestimmtheitskoeffizienten von 0.614 und einer ebenfalls fehlerfreien Klassifikation der Objekte.

3.3 Induktive Analyse

Im Rahmen einer induktiven Analyse der MD-Struktur unvollständiger Daten- und Distanzmatrizen sollen zuvor formulierte Hypothesen bezüglich der fehlenden Werte mittels statistischer Testverfahren überprüft werden. Dabei betreffen die in diesem Zusammenhang relevanten Hypothesen das Vorliegen von Häufungen fehlender Daten sowie von unsystematischen Ausfallmechanismen. Entsprechend werden im folgenden die zur Überprüfung der jeweiligen Hypothesen geeigneten Testverfahren vorgestellt.

3.3.1 Tests auf Häufungen fehlender Daten

Im Zusammenhang mit dem Vorliegen von Häufungen fehlender Daten in einer unvollständigen Daten- oder Distanzmatrix sind grundsätzlich zwei unterschiedliche Fragestellungen von Bedeutung. Zum einen ist von Interesse, inwieweit generell eine überhöhte Zahl fehlender Werte vorliegt und zum anderen stellt sich die Frage, ob merkmals- oder objektspezifische Häufungen fehlender Daten existieren.

Die Frage nach dem Vorliegen einer generell überhöhten Anzahl fehlender Werte kann dahingehend präzisiert werden, daß fehlende Daten als seltene Ereignisse betrachtet werden und damit als hypothetisches Verteilungsmodell die Poisson-Verteilung zugrundegelegt wird (vgl. *Lösel, Wüstendörfer, 1974, S. 344*). Für den Fall einer unvollständigen Datenmatrix ist dann zu überprüfen, ob die Verteilung der gemäß (3.3) bzw. (3.4) bestimmten Werte $v_{i\bullet}^{mis}(i=1,\ldots,n)$ bzw. $v_{\bullet k}^{mis}(k=1,\ldots,m)$, also die Verteilung der Anzahl fehlender Daten der einzelnen Objekte bzw. Merkmale, jeweils zu dem Verteilungstyp der Poisson-Verteilung gehört. Die Hypothesen werden also folgendermaßen formuliert:

$$\begin{array}{|l|}\hline H_0\text{: die Werte } v_{i\bullet}^{mis}(i=1,\ldots,n) \text{ entstammen einer } P(\lambda)\text{-Verteilung} \\ H_1\text{: die Werte } v_{i\bullet}^{mis}(i=1,\ldots,n) \text{ entstammen keiner } P(\lambda)\text{-Verteilung} \\ \hline\end{array} \tag{3.42}$$

$$\begin{array}{|l|}\hline H_0\text{: die Werte } v_{\bullet k}^{mis}(k=1,\ldots,m) \text{ entstammen einer } P(\lambda)\text{-Verteilung} \\ H_1\text{: die Werte } v_{\bullet k}^{mis}(k=1,\ldots,m) \text{ entstammen keiner } P(\lambda)\text{-Verteilung} \\ \hline\end{array} \tag{3.43}$$

Entsprechend sind bei Vorliegen einer unvollständigen, unmittelbar erhobenen Distanzmatrix die folgenden Hypothesen zu überprüfen:

$$\begin{array}{|l|}\hline H_0\text{: die Werte } w_i^{mis}(i=1,\ldots,n) \text{ entstammen einer } P(\lambda)\text{-Verteilung} \\ H_1\text{: die Werte } w_i^{mis}(i=1,\ldots,n) \text{ entstammen keiner } P(\lambda)\text{-Verteilung} \\ \hline\end{array} \tag{3.44}$$

Eine Methode zum Testen der in (3.42) bis (3.44) aufgestellten Hypothesen stellt der χ^2-Anpassungstest dar (vgl. z.B. *Bamberg, Baur, 1993, S. 198-202*). Da in allen drei Fällen

das Vorliegen des Verteilungstyps der Poisson-Verteilung überprüft werden soll, muß zunächst ein Maximum-Likelihood-Schätzwert $\hat{\lambda}$ für den Parameter λ bestimmt werden, der sich jeweils als arithmetisches Mittel der Werte $v_{i\bullet}^{mis}(i=1,\ldots,n)$ bzw. $v_{\bullet k}^{mis}(k=1,\ldots,m)$ bzw. $w_{i\bullet}^{mis}(i=1,\ldots,n)$ ergibt.[2] Die Grenzverteilung der resultierenden Testfunktion ist dann zwar keine χ^2-Verteilung mehr, jedoch liegt das zu dieser Grenzverteilung gehörende $(1-\alpha)$-Fraktil zwischen den entsprechenden Fraktilen der $\chi^2(\beta-1)$- und der $\chi^2(\beta-2)$-Verteilung. Dabei bezeichnen α das Signifikanzniveau und β die Anzahl der aneinander angrenzenden Intervalle, in die alle jeweils zu untersuchenden Werte zur Bestimmung des Testfunktionswerts eingeteilt werden (vgl. *Albrecht, 1980, S. 152-154*).

Beispiel:

Für die Datenmatrix des Anhangs A (Fall 1) soll untersucht werden, inwiefern eine generell überhöhte Anzahl fehlender Daten vorliegt. Verwendet man dazu die Anzahl der fehlenden Werte der einzelnen Objekte, die gemäß (3.3) berechnet werden, dann sind die Hypothesen gemäß (3.42) zu testen. Als Maximum-Likelihood-Schätzwert für den Parameter λ ergibt sich

$$\hat{\lambda}=\frac{1}{15}\sum_{i=1}^{15}v_{i\bullet}^{mis}=\frac{13}{15}.$$

Damit erhält man in jedem der folgenden Bereiche B_t $(t=1,2,3)$ die jeweils angegebene beobachtete Anzahl h_t sowie erwartete Anzahl $n\cdot p_t$ der in B_t liegenden Werte $v_{i\bullet}^{mis}(i=1,\ldots,n)$, wobei p_t die Wahrscheinlichkeit dafür ist, daß ein Wert $v_{i\bullet}^{mis}$ unter der Annahme einer $P(\frac{13}{15})$-Verteilung in den Bereich B_t fällt:

t	1	2	3
B_t	{0}	{1}	{2,3,...}
h_t	7	4	4
$n\cdot p_t$	15·0.4204 = 6.306	15·0.3643 = 5.4645	15·0.2153 = 3.2295

Als Testfunktionswert T ergibt sich damit ein Wert von $T=\sum_{t=1}^{3}\frac{(h_t-n\cdot p_t)^2}{n\cdot p_t}=0.6527$.

---

2 Die Likelihood-Funktion einer einfachen Stichprobe vom Umfang n einer $P(\lambda)$-verteilten Grundgesamtheit $(\lambda>0)$ ergibt sich gemäß

$$f(x_1,\ldots,x_n|\lambda)=\frac{\lambda^{x_1}}{x_1!}\cdot e^{-\lambda}\cdot\ldots\cdot\frac{\lambda^{x_n}}{x_n!}\cdot e^{-\lambda}=\frac{\lambda^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}x_i!}\cdot e^{-n\cdot\lambda}.$$

Die Ableitung der logarithmierten Likelihood-Funktion erhält man mit

$$\frac{\partial}{\partial\lambda}\left[\ln f(x_1,\ldots,x_n|\lambda)\right]=\frac{\partial}{\partial\lambda}\left[\sum_{i=1}^{n}x_i\cdot\ln\lambda-\ln\prod_{i=1}^{n}x_i!-n\cdot\lambda\right]=\frac{1}{\lambda}\cdot\sum_{i=1}^{n}x_i-n.$$

Durch Nullsetzen ergibt sich als Maximum-Likelihood-Schätzwert das Stichprobenmittel, d.h.

$$\hat{\lambda}=\frac{1}{n}\sum_{i=1}^{n}x_i,$$

jedoch muß dieses positiv sein, da nur dann die zweite Ableitung der Likelihood-Funktion negativ wird.

Bei einem Signifikanzniveau von $\alpha = 0.05$ kann die Nullhypothese nicht verworfen werden, da der empirische Testfunktionswert T kleiner als das zur theoretischen Grenzverteilung von T gehörende 0.95-Fraktil, das zwischen den 0.95-Fraktilen der $\chi^2(1)$- und $\chi^2(2)$-Verteilung, also zwischen 3.84 und 5.99 liegt, ist. Anzumerken ist, daß in diesem Beispiel die Abschätzung des zur Grenzverteilung von T gehörenden α-Fraktils ungenau sein kann, da in den festgelegten Bereichen teilweise eine zu geringe Anzahl beobachteter bzw. erwarteter Häufigkeiten vorliegt. (Grundsätzlich muß für eine hinreichend genaue Abschätzung des theoretischen Testfunktionswerts $h_t \geq 5$ oder $n \cdot p_t \geq 5 \; \forall t$ gelten.) Wird jedoch die Anzahl der Bereiche in der Art verkleinert, daß die Bereiche B_2 und B_3 zusammengefaßt werden, dann ist eine Abschätzung des theoretischen Testfunktionswerts mittels der entsprechenden α-Fraktile der χ^2-Verteilung nicht mehr möglich. Eine statistische Überprüfung der gemäß (3.43) bzw. (3.44) formulierten Hypothesen für die Daten des Anhangs A bzw. B ist aufgrund der zu geringen Merkmals- bzw. Objektanzahl nicht durchführbar.

Zur Untersuchung merkmals- und objektspezifischer Häufungen der fehlenden Werte einer unvollständigen Datenmatrix bietet sich das Zugrundelegen der folgenden, allerdings von der Stichprobe abhängigen hypothetischen Verteilung an: Die möglichen positiven Ausprägungen der Anzahl fehlender Daten der einzelnen Merkmale bzw. Objekte, die zwischen dem jeweiligen Minimum und Maximum liegen, werden als gleichwahrscheinlich betrachtet. Zu Überprüfen sind damit die Hypothesen gemäß (3.45) und (3.46).

H_0: die Werte $v_{i\bullet}^{mis} \neq 0 \; (i = 1,\ldots,n)$ entstammen einer Verteilung mit

$$f\left(v_{i\bullet}^{mis}\right) = \begin{cases} \dfrac{1}{1+\max\limits_{i\in N} v_{i\bullet}^{mis} - \min\limits_{i\in N: v_{i\bullet}^{mis}\neq 0} v_{i\bullet}^{mis}} & \text{für } v_{i\bullet}^{mis} \in \left\{ \min\limits_{i\in N: v_{i\bullet}^{mis}\neq 0} v_{i\bullet}^{mis}, \ldots, \max\limits_{i\in N} v_{i\bullet}^{mis} \right\} \\ 0 & \text{sonst} \end{cases} \tag{3.45}$$

H_1: die Werte $v_{i\bullet}^{mis} \neq 0 \; (i = 1,\ldots,n)$ entstammen nicht dieser Verteilung

H_0: die Werte $v_{\bullet k}^{mis} \neq 0 \; (k = 1,\ldots,m)$ entstammen einer Verteilung mit

$$f\left(v_{\bullet k}^{mis}\right) = \begin{cases} \dfrac{1}{1+\max\limits_{k\in M} v_{\bullet k}^{mis} - \min\limits_{k\in M: v_{\bullet k}^{mis}\neq 0} v_{\bullet k}^{mis}} & \text{für } v_{\bullet k}^{mis} \in \left\{ \min\limits_{k\in M: v_{\bullet k}^{mis}\neq 0} v_{\bullet k}^{mis}, \ldots, \max\limits_{k\in M} v_{\bullet k}^{mis} \right\} \\ 0 & \text{sonst} \end{cases} \tag{3.46}$$

H_1: die Werte $v_{\bullet k}^{mis} \neq 0 \; (k = 1,\ldots,m)$ entstammen nicht dieser Verteilung

Analog sind im Fall einer unvollständigen, unmittelbar erhobenen Distanzmatrix lediglich die objektspezifischen Häufungen mittels der Hypothesen

H_0: die Werte $w_i^{mis} \neq 0 \; (i = 1,\ldots,n)$ entstammen einer Verteilung mit

$$f\left(w_i^{mis}\right) = \begin{cases} \dfrac{1}{1+\max\limits_{i\in N} w_i^{mis} - \min\limits_{i\in N: w_i^{mis}\neq 0} w_i^{mis}} & \text{für } w_i^{mis} \in \left\{ \min\limits_{i\in N: w_i^{mis}\neq 0} w_i^{mis}, \ldots, \max\limits_{i\in N} w_i^{mis} \right\} \\ 0 & \text{sonst} \end{cases} \tag{3.47}$$

H_1: die Werte $w_i^{mis} \neq 0 \; (i = 1,\ldots,n)$ entstammen nicht dieser Verteilung

zu untersuchen. Als Testmethode kann wiederum der χ^2-Anpassungstest herangezogen werden. Da die in den Verteilungen verwendeten Parameter jeweils Maximum-Likelihood-Schätzwerte darstellen,[3] liegt das zur Grenzverteilung des resultierenden Testfunktionswerts gehörende $(1-\alpha)$-Fraktil zwischen den entsprechenden Fraktilen der $\chi^2(\beta-1)$- und der $\chi^2(\beta-2)$-Verteilung. Dabei werden wiederum mit α das Signifikanzniveau und mit β die Anzahl der aneinander angrenzenden Intervalle, in die alle jeweils zu untersuchenden Werte zur Bestimmung des Testfunktionswerts eingeteilt werden, bezeichnet.

Beispiel:

Betrachtet man die Datenmatrix des Anhangs A (Fall 1), dann sind im Rahmen einer Untersuchung objektspezifischer Häufungen der fehlender Daten die Hypothesen (3.45) und damit eine hypothetische Verteilung der Werte $v_{i\bullet}^{mis}(i=1,\ldots,n)$ mit der Wahrscheinlichkeitsfunktion

$$f\left(v_{i\bullet}^{mis}\right)=\begin{cases}\dfrac{1}{1+3-1}=\dfrac{1}{3} & \text{für } v_{i\bullet}^{mis}\in\{1,2,3\}\\ 0 & \text{sonst}\end{cases}$$

zu testen. Die Durchführung eines χ^2-Anpassungstests ist jedoch aufgrund der zu geringen Anzahl vorliegender Werte $v_{i\bullet}^{mis}\neq 0$ nicht möglich. Die Hypothesen (3.46) und, bei Betrachtung der Distanzmatrix des Anhangs B, die Hypothesen (3.47) können ebenfalls aufgrund einer jeweils zu geringen Anzahl vorliegender Werte $v_{\bullet k}^{mis}\neq 0$ bzw. $w_i^{mis}\neq 0$ nicht überprüft werden.

Wie aus den beiden vorherigen Beispielen ersichtlich wird, haben die in diesem Abschnitt dargestellten Möglichkeiten einer Untersuchung von Häufungen fehlender Daten im Rahmen einer induktiven Analyse entscheidende Nachteile. Neben der grundsätzlich benötigten hohen Anzahl an Objekten bzw. Merkmalen muß vor allem für die Hypothesen (3.45) bis (3.47) auch eine hohe Anzahl an Objekten bzw. Merkmalen mit fehlenden Daten vorliegen. Damit wird die praktische Anwendung der dargestellten Testansätze stark eingeschränkt.

---

[3] Die Likelihood-Funktion einer einfachen Stichprobe vom Umfang n, die aus einer Grundgesamtheit stammt, deren Verteilung mit der Wahrscheinlichkeitsfunktion

$$f(x)=\begin{cases}\dfrac{1}{1+b-a} & \text{für } x\in\{a,\ldots,b\}\\ 0 & \text{sonst}\end{cases}$$

gegeben ist, ergibt sich gemäß

$$f(x_1,\ldots,x_n|\,a,b)=\frac{1}{1+b-a}\cdot\ldots\cdot\frac{1}{1+b-a}=\frac{1}{(1+b-a)^n}.$$

Die Maximierung der Likelihood-Funktion bedeutet, daß die Differenz $(b-a)$ minimal werden muß. Dies wird durch $a=\min_i x_i$ und $b=\max_i x_i$ erreicht.

3.3.2 Tests auf unsystematische Ausfallmechanismen

Mit den hier betrachteten Testverfahren soll das Vorliegen unsystematischer Ausfallmechanismen überprüft werden. Die Hypothesen lassen sich dabei in der folgenden allgemeinen Form formulieren:

$$\boxed{\begin{array}{l} H_0\text{: die Daten fehlen zufällig} \\ H_1\text{: die Daten fehlen systematisch} \end{array}} \tag{3.48}$$

Grundsätzlich ist festzuhalten, daß alle Tests zur Überprüfung der Hypothesen gemäß (3.48) nie das zufällige Fehlen von Daten bestätigen, sondern lediglich bestimmte Formen eines zufälligen Fehlens ausschließen können (vgl. *Frane, 1978, S. 27*). Falls ein derartiger Ausschluß nicht möglich ist, d.h. H_0 nicht abgelehnt werden kann, stellt dies allenfalls eine notwendige Bedingung zur Akzeptierung eines unsystematischen Fehlens der Daten dar. Die einzelnen Tests zielen also auf einzelne Aspekte eines unsystematischen Fehlens der Daten und damit auf bestimmte Ausfallmechanismen ab. Dabei können die in der Tabelle 3.6 angegebenen Zielsetzungen einschließlich der jeweiligen Zielrichtung, d.h. welche Eigenschaft der Daten im Hinblick auf den Ausfallmechanismus untersucht wird, des jeweiligen Ausgangspunktes und der geeigneten Testverfahren unterschieden werden. Dabei ist anzumerken, daß nur in der Gesamtbetrachtung der Testergebnisse aller Testansätze eine sinnvolle Annahme über den Ausfallmechanismus, der den Daten zugrundeliegt, möglich ist.

Zielsetzung	**Zielrichtung**	**Ausgangspunkt**	**Testverfahren**
Untersuchung der Abhängigkeit der MD von den an sich unbekannten Realisierungen dieser Werte	MAR, OAR	Daten- bzw. Distanzmatrix	Anpassungstests, parametrische Einstichprobentests
Untersuchung der Abhängigkeit der MD vom Fehlen der Daten bei anderen Merkmalen bzw. Objekten	MAR, OAR	Indikatormatrix	Test nach *Kim und Curry*, Tests bezüglich der Korrelationskoeffizienten
Untersuchung der Abhängigkeit der MD von den vorhandenen Ausprägungen bei anderen Merkmalen bzw. Objekten	OAR	Daten- und Indikatormatrix	Tests auf Lokationsunterschiede bzw. Unabhängigkeit, Test nach *Little*

Tabelle 3.6: Überblick der Testansätze und -verfahren

Da sich die Tests lediglich auf die vorhandenen Daten sowie auf die Indikatormatrix beziehen können, ergeben sich einige Probleme. Zum einen können im Fall einer unvollständigen Datenmatrix die fehlenden Werte von den fehlenden Ausprägungen bei anderen Merkmalen bzw. Objekten abhängen. Zum anderen ist die Abhängigkeit der fehlen-

den Daten von den an sich unbekannten Realisierungen dieser fehlenden Werte nicht direkt überprüfbar. Somit kann das Vorliegen der Eigenschaft MAR anhand der vorhandenen Daten ohnehin nicht vollständig getestet werden. Eine Lösung dieses Problems ist beispielsweise durch eine sogenannte Follow-up-Erhebung möglich (vgl. *Toutenburg, 1992, S. 202-203*). Jedoch müßte es sich bei dieser Nacherhebung von zunächst fehlenden Daten um eine zufällige Stichprobe handeln, die entweder ohne fehlende Daten ist oder bei der die Daten zufällig fehlen. Es wird also deutlich, daß ein systematisches Fehlen der Daten nie völlig ausgeschlossen werden kann. Gerade deshalb ist nicht ausschließlich die Frage nach dem Vorliegen eines systematischen Ausfallmechanismus von Bedeutung, sondern auch die Frage nach der Stärke eines derartigen Effekts. So müssen selbst im Fall eines signifikanten Ergebnisses für das Vorliegen systematisch fehlender Daten die resultierenden Analyseergebnisse nicht völlig in Frage gestellt werden (vgl. *Schnell, 1986, S. 12-13*).

Für die ausschließlich auf der Daten- bzw. Distanzmatrix basierenden Tests, mit denen die Abhängigkeit der fehlenden Daten von den unbekannten Realisierungen dieser Werte analysiert werden soll, müssen externe Informationen herangezogen werden. Eine derartige externe Information stellt beispielsweise die Verteilung der Ausgangsstichprobe dar. Damit werden diese Testansätze in ihrer praktischen Anwendung stark eingeschränkt. Die Tests zur Untersuchung der Abhängigkeit der fehlenden Werte vom Fehlen der Daten bei anderen Merkmalen bzw. Objekten sind für eine praktische Anwendung erheblich besser geeignet, da sie lediglich die in jedem Fall vollständig vorliegende Indikatormatrix heranziehen. Die auf der Daten- und Indikatormatrix basierenden Tests, mit denen die Abhängigkeit der fehlenden Daten von den vorhandenen Ausprägungen bei anderen Merkmalen bzw. Objekten untersucht werden soll, können dann zu Einschränkungen in der Anwendbarkeit führen, sobald die fehlenden Daten in den Tests zu berücksichtigen sind. Im Rahmen der nun folgenden Darstellung der einzelnen Testansätze und -verfahren wird diese hier kurz skizzierte Problematik an entsprechender Stelle ausführlich diskutiert.

Die induktive Analyse der Abhängigkeit der MD von den an sich unbekannten Realisierungen dieser Werte kann mittels **Anpassungstests** durchgeführt werden. Im Fall einer Datenmatrix ist dabei zu überprüfen, inwieweit die für ein Merkmal vorliegenden Daten einer hypothetischen Verteilung genügen. Falls es sich bei der ursprünglichen Stichprobe um eine zufällige Stichprobe aus der Grundgesamtheit handelt, kann als hypothetische Verteilung die der Grundgesamtheit herangezogen werden. Falls es sich um keine zufällige Stichprobe handelt, müßte die Verteilung dieser Stichprobe als hypothetische Verteilung bekannt sein. Die Daten können dann als zufällig fehlend im Sinne der Zielsetzung dieses Testansatzes angesehen werden, wenn die vorhandenen Daten der hypothetischen Verteilung genügen. Im umgekehrten Fall kann die Annahme MAR für

die Daten nicht akzeptiert werden. Bei Vorliegen einer unvollständigen, unmittelbar erhobenen Distanzmatrix ist zu überprüfen, inwieweit die insgesamt vorhandenen paarweisen Distanzen der Verteilung der ursprünglich erhobenen Distanzen entsprechen. Dabei ergibt sich die hypothetische Verteilung analog zu den Ausführungen im Fall einer Datenmatrix. Auch in diesem Fall ist die Annahme MAR für die Daten nicht tragbar, falls die empirische und die hypothetische Verteilung der paarweisen Distanzen signifikant voneinander abweichen. Die Tatsache, daß keine signifikante Abweichung zwischen empirischer und hypothetischer Verteilung vorliegt, stellt wiederum nur eine notwendige Bedingung zur Akzeptierung eines unsystematischen Fehlens der Daten dar. Als mögliche Testverfahren seien an dieser Stelle beispielsweise der χ^2-Anpassungstest oder der Kolmogoroff-Smirnoff-Test erwähnt.

Beispiel:

Für die Datenmatrix des Anhangs A soll zunächst für den Fall des unsystematischen Ausfallmechanismus (Fall 1) untersucht werden, inwiefern ein zufälliges Fehlen der Daten beim Merkmal Preisniveau (1) im Hinblick auf die Abhängigkeit der MD von den an sich unbekannten Realisierungen dieser Werte vorliegt. Als hypothetische Verteilung wird die als bekannt vorausgesetzte Verteilung der Ausgangsstichprobe, d.h. die Verteilung der Ausprägungen aller 15 Objekte, verwendet. Die zugehörige Wahrscheinlichkeitsfunktion besitzt die folgende Gestalt:

$$f(a_{i1}) = \begin{cases} \frac{2}{15} & \text{für } a_{i1} = 1 \text{ (niedriges Preisniveau)} \\ \frac{6}{15} & \text{für } a_{i1} = 2 \text{ (mittleres Preisniveau)} \\ \frac{2}{15} & \text{für } a_{i1} = 3 \text{ (gehobenes Preisniveau)} \\ \frac{6}{15} & \text{für } a_{i1} = 4 \text{ (hohes Preisniveau)} \end{cases}.$$

Damit erhält man in jedem der Bereiche $B_t (t = 1,2)$ die jeweils in der nachfolgenden Tabelle angegebene beobachtete Anzahl h_t sowie erwartete Anzahl $n \cdot p_t$ der in B_t liegenden vorhandenen Ausprägungen des Merkmals Preisniveau. Dabei bezeichnet p_t die Wahrscheinlichkeit, daß eine Ausprägung unter der Annahme der hypothetischen Verteilung in den Bereich B_t fällt. Eine feinere Einteilung der Bereiche ist aufgrund der Voraussetzung $h_t \geq 5$ oder auch $n \cdot p_t \geq 5$ nicht möglich.

t	1	2
B_t	$\{1,2\}$	$\{3,4\}$
h_t	7	6
$n \cdot p_t$	$13 \cdot \frac{8}{15} = 6\frac{14}{15}$	$13 \cdot \frac{7}{15} = 6\frac{1}{15}$

Als Testfunktionswert T ergibt sich damit ein Wert von $T = \sum_{t=1}^{2} \frac{(h_t - n \cdot p_t)^2}{n \cdot p_t} = 0.0014$.

Bei einem Signifikanzniveau von $\alpha = 0.05$ kann die Nullhypothese nicht verworfen werden, da der empirische Testfunktionswert T kleiner als das zur theoretischen Grenzverteilung von T gehörende 0.95-Fraktil der $\chi^2(1)$-Verteilung mit 3.84 ist. Damit kann die Annahme der Eigenschaft MCAR für die Daten zunächst

nicht verworfen werden. Dies stellt jedoch lediglich eine notwendige und keine hinreichende Bedingung zur Akzeptierung eines zufälligen Fehlens der Daten dar.

Als nächstes soll die im Fall 2 der Datenmatrix des Anhangs A berücksichtigte Abhängigkeit der fehlenden Daten von den tatsächlichen Realisierungen dieser Werte beim Merkmal Multivariate Verfahren untersucht werden. Als hypothetische Verteilung wird wiederum die Verteilung der Ausgangsstichprobe verwendet, wobei eine N(35.45;20.25)-Verteilung unterstellt wird. Analog zum Fall zufällig fehlender Daten ergibt sich für die vorliegenden Ausprägungen beim Merkmal Multivariate Verfahren die folgende Arbeitstabelle:

t	1	2
B_t	$\langle -\infty;35.45]$	$\langle 35.45;\infty\rangle$
h_t	9	1
$n \cdot p_t$	5	5

Als Testfunktionswert erhält man einen Wert von $T = 6.40$. Bei einem Signifikanzniveau von $\alpha = 0.05$ wird die Nullhypothese verworfen, da in diesem Fall der empirische Testfunktionswert T größer als das zur theoretischen Grenzverteilung von T gehörende 0.95-Fraktil der $\chi^2(1)$-Verteilung mit 3.84 ist. Damit wird das Vorliegen eines systematischen Ausfallmechanismus bestätigt.

Eine entsprechende Untersuchung soll nun für die Distanzmatrix des Anhangs B vorgenommen werden. Dabei wird die folgende Dichtefunktion für die ursprünglich 45 zu erhebenden paarweisen Distanzen unterstellt:

$$f(d_{ij}) = \begin{cases} 0.05 \cdot d_{ij} & \text{für } d_{ij} \in [0;4] \\ 0.15 \cdot d_{ij} - 0.4 & \text{für } d_{ij} \in \langle 4;5] \\ 1.8 - 0.29 \cdot d_{ij} & \text{für } d_{ij} \in \langle 5;6] \\ 0.015 - 0.015 \cdot d_{ij} & \text{für } d_{ij} \in \langle 6;10] \\ 0 & \text{sonst} \end{cases}.$$

Damit erhält man zunächst für den Fall zufällig fehlender Distanzen (Fall 1) die in der folgenden Tabelle für jedes Intervall B_t ($t = 1,2,\ldots,7$) angegebene beobachtete Anzahl h_t sowie erwartete Anzahl $n \cdot p_t$ der in B_t liegenden vorhandenen paarweisen Distanzen. Mit p_t wird wiederum die Wahrscheinlichkeit bezeichnet, daß eine paarweise Distanz unter der Annahme der hypothetischen Verteilung in das Intervall B_t fällt.

t	1	2	3	4	5	6	7
B_t	$\langle -\infty;2.39]$	$\langle 2.39;3.38]$	$\langle 3.38;4.13]$	$\langle 4.13;4.69]$	$\langle 4.69;5.12]$	$\langle 5.12;5.76]$	$\langle 5.76;\infty\rangle$
h_t	6	4	6	5	3	6	6
$n \cdot p_t$	5.14	5.14	5.14	5.14	5.14	5.14	5.14

Als Testfunktionswert ergibt sich damit ein Wert von $T = 1.72$. Bei einem Signifikanzniveau von $\alpha = 0.05$ und dem 0.95-Fraktil der $\chi^2(6)$-Verteilung mit 12.59 kann eine Abhängigkeit der fehlenden paarweisen Distanzen von den an sich unbekannten Realisierungen dieser Werte und damit ein systematisches Fehlen der Daten nicht bestätigt werden.

Analog zum Fall zufällig fehlender Distanzen ergibt sich bei Betrachtung der im Fall 2 des Anhangs B berücksichtigten Abhängigkeit der fehlenden Distanzen von den tatsächlichen Realisierungen dieser Werte die folgende Arbeitstabelle:

t	1	2	3	4	5	6	7
B_t	$\langle -\infty; 2.39]$	$\langle 2.39; 3.38]$	$\langle 3.38; 4.13]$	$\langle 4.13; 4.69]$	$\langle 4.69; 5.12]$	$\langle 5.12; 5.76]$	$\langle 5.76; \infty\rangle$
h_t	6	6	8	6	4	6	0
$n \cdot p_t$	5.14	5.14	5.14	5.14	5.14	5.14	5.14

Als Testfunktionswert erhält man hier einen Wert von $T = 6.99$. Der Testfunktionswert ist zwar deutlich größer als im Fall zufällig fehlender Daten, jedoch kann bei einem Signifikanzniveau von $\alpha = 0.05$ die Nullhypothese nicht verworfen werden. Damit zeigt sich, daß die Anzahl beobachteter Distanzen im Intervall B_7 mit einem Wert von Null, die gemäß dem zugrundegelegten Ausfallmechanismus resultiert, nicht ausreicht, die gesamte Verteilung zu verwerfen.

Anhand des Beispiels wird deutlich, daß ein Anpassungstest im Fall größerer Datenmengen mit relativ wenigen fehlenden Werten im allgemeinen nicht geeignet ist, die Abhängigkeit der fehlenden Daten von den tatsächlichen Realisierungen dieser Werte aufzuzeigen. Anstelle der Untersuchung der gesamten Verteilung des Datenmaterials bietet sich daher lediglich die Betrachtung einzelner Verteilungsparameter an, da diese eher signifikante Abweichungen aufweisen werden. Entsprechend sind **parametrische Einstichprobentests** durchzuführen, mit denen der Vergleich eines Parameters der vorliegenden Stichprobe mit dem der Ausgangsstichprobe möglich ist. Falls es sich bei der Ausgangsstichprobe wiederum um eine zufällige Stichprobe handelt, kann als Vergleichsparameter auch der entsprechende Parameter der Grundgesamtheit herangezogen werden. Als zu untersuchende Parameter bieten sich bei kardinal skalierten Daten unter anderem die Varianz und vor allem der Mittelwert an. Damit können der **χ^2-Test für die Varianz** sowie, je nach Verteilungsannahme, der **Einstichproben-Gaußtest**, der **Einstichproben-t-Test** und der **approximative Gaußtest** zur Anwendung kommen. Eine ausführliche Beschreibung dieser Testverfahren findet man beispielsweise bei *Bamberg und Baur (1993, S. 187-192)*. Im Fall ordinal skalierter Daten ist hingegen der **Einstichproben-Vorzeichentest** geeignet, mit dem ein Vergleich des Medians der vorhandenen Daten mit dem Median der Ausgangsstichprobe möglich ist (vgl. z.B. *Hartung, 1989, S. 242-243*).

Beispiel:

Zunächst soll für die Datenmatrix des Anhangs A der Fall zufällig fehlender Werte (Fall 1) betrachtet werden. Für die Merkmale Deskriptive Statistik, Spezialgebiete und Statistische Grafiken wird jeweils eine Normalverteilung mit bekannter Varianz unterstellt. Aus Gründen der Vereinfachung wird die Varianz der Ausgangsdaten verwendet. Für einen Vergleich des Mittelwerts der jeweils vorhandenen Daten mit

dem als bekannt vorausgesetzten Mittelwert der Ausgangsstichprobe ist der Einstichproben-Gaußtest anzuwenden. Es ergibt sich die nachfolgende Arbeitstabelle, wobei μ_0 und σ den Mittelwert und die Standardabweichung der Ausgangsstichprobe und $\bar{x}$ den Mittelwert der vorhandenen Daten für die einzelnen Merkmale bezeichnen. Als Signifikanzniveau wird $\alpha = 0.05$ gewählt.

Merkmal	Deskriptive Statistik	Spezialgebiete	Statistische Grafiken
μ_0	77.93	44.97	37.82
σ	11.82	24.76	23.55
$\bar{x}$	76.61	44.47	40.15
Testfunktionswert	-0,43	-0,08	0.38
Verwerfungsbereich	$\langle -\infty;-1.96\rangle \cup \langle 1.96;\infty\rangle$	$\langle -\infty;-1.96\rangle \cup \langle 1.96;\infty\rangle$	$\langle -\infty;-1.96\rangle \cup \langle 1.96;\infty\rangle$

Bei allen drei Merkmalen kann also die Nullhypothese, daß die Mittelwerte der vorhandenen Daten gleich den Mittelwerten der Ausgangsstichprobe sind, nicht verworfen werden. Dies spricht für ein zufälliges Fehlen der Daten. Führt man unter den gleichen Testvoraussetzungen den Einstichproben-Gaußtest für den im Anhang A beschriebenen Fall 2 durch, dann ergeben sich für das Merkmal Multivariate Verfahren ein Mittelwert und eine Standardabweichung der Ausgangsstichprobe von $\mu_0 = 35.45$ und $\sigma = 20.25$, ein Mittelwert der vorhandenen Daten von $\bar{x} = 22.27$ und ein Testfunktionswert von -2.06. Bei einem Signifikanzniveau von $\alpha = 0.05$ liegt der Testfunktionswert im Verwerfungsbereich $\langle -\infty;-1.96\rangle \cup \langle 1.96;\infty\rangle$, so daß die vorliegende Abhängigkeit der fehlenden Daten von den tatsächlichen Realisierungen dieser Werte zum Ausdruck kommt.

Für die Distanzmatrix des Anhangs B werden im folgenden die Fälle 1 und 2, also das zufällige sowie systematische Fehlen der paarweisen Distanzen, gleichzeitig betrachtet. Die Distanzen werden als kardinal skaliert angesehen und es wird eine beliebige Verteilung mit bekannter Varianz unterstellt, wobei die Varianz der 45 ursprünglich zu erhebenden paarweisen Distanzen verwendet wird. Damit kommt der approximative Gaußtest zur Anwendung und es ergibt sich die folgende Arbeitstabelle, in der μ_0 und σ den Mittelwert und die Standardabweichung der Ausgangsstichprobe und $\bar{x}$ den Mittelwert der vorhandenen, von den ursprünglich 45 zu erhebenden Distanzen bezeichnen:

Ausfallmechanismus	Fall 1	Fall 2
μ_0	4.39	4.39
σ	1.61	1.61
$\bar{x}$	4.40	3.80
Testfunktionswert	0.04	-2.20

Bei einem Signifikanzniveau von $\alpha = 0.05$ liegt der Testfunktionswert im Fall zufällig fehlender Distanzen außerhalb und im Fall der Abhängigkeit der fehlenden Distanzen von den tatsächlichen Realisierungen dieser Werte innerhalb des Verwerfungsbereichs $\langle -\infty;-1.96\rangle \cup \langle 1.96;\infty\rangle$. Diese Ergebnisse führen damit zu den richtigen Aussagen über den jeweils zugrundeliegenden Ausfallmechanismus.

Ein geeignetes Testverfahren des zweiten Testansatzes, also der Untersuchung der Abhängigkeit der MD vom Fehlen der Daten anderer Merkmale bzw. Objekte, ist der **Test nach *Kim und Curry*** (vgl. *Kim, Curry, 1977, S. 219*). Dieser Test basiert ausschließlich auf der Indikatormatrix und ist sowohl bei Vorliegen einer unvollständigen Datenmatrix wie auch bei Vorliegen einer unvollständigen, unmittelbar erhobenen Distanzmatrix anwendbar. Die zu prüfenden Hypothesen und der Testablauf sollen im folgenden lediglich für den Fall einer merkmalsweisen Untersuchung einer unvollständigen Datenmatrix aufgezeigt werden. Diese Darstellungen sind problemlos auf den Fall einer objektweisen Untersuchung der Datenmatrix sowie einer entsprechenden Untersuchung einer Distanzmatrix übertragbar. Die in (3.48) zunächst allgemein formulierten Hypothesen können nun gemäß der Zielsetzung dieses Testansatzes folgendermaßen präzisiert werden:

H_0: die Daten fehlen zufällig in dem Sinn, daß sie vom Fehlen der Daten anderer Merkmale unabhängig sind
H_1: die Daten fehlen systematisch (3.49)

Dabei ist wiederum anzumerken, daß bei einer Ablehnung der Nullhypothese die Annahme der Eigenschaften MAR und OAR und damit auch der Eigenschaft MCAR für die Daten nicht aufrecht erhalten werden kann, während die Bestätigung der Nullhypothese lediglich eine notwendige Bedingung für die Akzeptierung eines zufälligen Fehlens der Daten darstellt. Um die Vorgehensweise des Tests nach *Kim und Curry* zur Überprüfung der in (3.49) formulierten Hypothesen nachvollziehen zu können, erfolgt in Abbildung 3.19 eine formale Darstellung. Der Wert h_k bezeichnet dabei die tatsächliche Anzahl der Objekte, die ausschließlich bei Merkmal k eine fehlende Ausprägung besitzen, während die Werte h^{nom} bzw. h^{mom} die Anzahl der Objekte angeben, bei denen keine bzw. mindestens zwei Merkmalsausprägungen fehlen. Entsprechend symbolisieren die mit einer Schlange gekennzeichneten Werte die jeweils zugehörige, erwartete Anzahl an Objekten. Der Testfunktionswert z ergibt sich dann aus der Summe der relativen quadratischen Abweichungen zwischen den tatsächlichen und den erwarteten Häufigkeiten. Vom Grundprinzip entspricht der Test nach *Kim und Curry* damit dem χ^2-Test.

Gemäß *Stewart (1982, S. 395)* führt der Test nach *Kim und Curry* im Hinblick auf die in der Tabelle 3.6 angegebene Zielsetzung zu zufriedenstellenden Ergebnissen, wenn die Anzahl der fehlenden Daten sowie, im Fall der hier in erster Linie betrachteten merkmalsweisen Untersuchung einer unvollständigen Datenmatrix, die Anzahl der in der Datenmatrix vorliegenden Merkmale hoch ist. Darüber hinaus impliziert eine Erhöhung der Objektanzahl bei sonst gleichbleibender Struktur der Datenmatrix eine entsprechende Erhöhung des Testfunktionswerts. Wird also beispielsweise die Objektanzahl bei sonst gleichbleibender Struktur der Datenmatrix verdoppelt, dann verdoppelt sich auch

der Testfunktionswert. Dies bedeutet, daß eine für die vorliegende Objektmenge zunächst nicht signifikante Abhängigkeit der fehlenden Werte vom Fehlen der Daten bei anderen Merkmalen dann als systematisches Fehlen bestätigt werden kann, wenn eine entsprechend große Anzahl weiterer Objekte mit einem in der Gesamtbetrachtung identischen MD-Muster zur Objektmenge hinzugefügt wird.

Schritt 1: Festlegung eines Signifikanzniveaus α.

Schritt 2: Ermittlung des Testfunktionswertes z gemäß

$$z = \frac{\left(h^{nom} - \tilde{h}^{nom}\right)^2}{\tilde{h}^{nom}} + \frac{\left(h^{mom} - \tilde{h}^{mom}\right)^2}{\tilde{h}^{mom}} + \sum_{k \in M^{mis}} \frac{\left(h_k - \tilde{h}_k\right)^2}{\tilde{h}_k}$$

mit $M^{mis} = \left\{k \in M : v_{\bullet k}^{ind} = 0\right\}$,

$$h^{nom} = \sum_{i=1}^{n} v_{i\bullet}^{ind}, \quad \tilde{h}^{nom} = n \cdot \prod_{k \in M^{mis}} \tilde{v}_{\bullet k}^{obs},$$

$$h^{mom} = n - h^{nom} - \sum_{k \in M^{mis}} h_k, \quad \tilde{h}^{mom} = n - \tilde{h}^{nom} - \sum_{k \in M^{mis}} \tilde{h}_k,$$

$$h_k = \left| \left\{ i \in N : v_{ik} = 0 \wedge v_{il} = 1 \,\forall\, l \in M, l \neq k \right\} \right| ; k \in M^{mis},$$

$$\tilde{h}_k = n \cdot \tilde{v}_{\bullet k}^{mis} \cdot \prod_{l \in M^{mis},\, l \neq k} \tilde{v}_{\bullet l}^{obs} ; k \in M^{mis}.$$

Schritt 3: Festlegung des Verwerfungsbereichs $B = \langle x_{1-\alpha}; \infty \rangle$, wobei der Fraktilswert aus der χ^2-Verteilung mit $\left(\left|M^{mis}\right| + 1\right)$ Freiheitsgraden stammt.

Schritt 4: H_0 wird genau dann abgelehnt, wenn $z \in B$ gilt.

Abbildung 3.19: Test nach Kim und Curry

Grundsätzlich wäre an eine Verfeinerung des Testfunktionswerts dahingehend zu denken, daß die Anzahl der Objekte mit mindestens zwei fehlenden Ausprägungen weiter aufgeteilt wird in die Anzahl der Objekte mit jeweils genau zwei fehlenden Ausprägungen, genau drei fehlenden Ausprägungen usw.. Dies ist jedoch nicht zweckmäßig, da die hier dargestellte Teststatistik bei größeren Datenmatrizen im allgemeinen zu zufriedenstellenden Ergebnissen führt und sich bei einer kleineren Datenmatrix die Problematik stellt, daß unter Umständen die erwarteten Häufigkeiten nicht mehr größer gleich fünf sind. In diesem Fall ist dann die Verteilung des Testfunktionswerts nicht mehr hinreichend genau eine χ^2-Verteilung.

Beispiel:

Für die Datenmatrix des Anhangs A (Fall 1) sollen die gemäß (3.49) formulierten Hypothesen überprüft werden. Zur Durchführung des Tests nach *Kim und Curry* kann die folgende Arbeitstabelle aufgestellt werden:

Merkmal k	1	2	3	4	7	9
h_k	1	0	1	0	1	1
$\tilde{h}_k$	0.90	0.42	0.90	1.46	1.46	0.90

$h^{nom} = 7$	$h^{mom} = 4$
$\tilde{h}^{nom} = 5.83$	$\tilde{h}^{mom} = 3.13$

Damit ergibt sich ein Testfunktionswert von $z = 2.52$. Bei einem Signifikanzniveau von $\alpha = 0.05$ erhält man den Verwerfungsbereich $B = \langle 14.07; \infty \rangle$. Die Nullhypothese kann also nicht abgelehnt werden.

Im Fall einer objektweisen Betrachtung der Datenmatrix des Anhangs A (Fall 1) ergibt sich, bei entsprechender Modifikation der Hypothesen und des Testablaufs, die folgende Arbeitstabelle zur Berechnung des Testfunktionswerts:

Objekt i	2	4	9	10	11	12	14	15
h_i	0	1	0	0	0	0	0	0
$\tilde{h}_i$	0.50	0.88	0.22	0.22	0.22	0.50	0.50	0.22

$h^{nom} = 3$	$h^{mom} = 5$
$\tilde{h}^{nom} = 1.76$	$\tilde{h}^{mom} = 3.98$

Als Testfunktionswert erhält man $z = 3.55$. Bei einem Signifikanzniveau von $\alpha = 0.05$ kann die Nullhypothese aufgrund des Verwerfungsbereich $B = \langle 16.92; \infty \rangle$ nicht abgelehnt werden.

Betrachtet man den für die Datenmatrix im Anhang A beschriebenen Fall 3, der die wechselseitige Abhängigkeit der fehlenden Ausprägungen bei den Merkmalen Businessgrafiken (8) und Statistische Grafiken (9) berücksichtigt, dann erhält man die folgende Arbeitstabelle:

Merkmal k	8	9
h_k	0	0
$\tilde{h}_k$	3.33	3.33

$h^{nom} = 10$	$h^{mom} = 5$
$\tilde{h}^{nom} = 6.67$	$\tilde{h}^{mom} = 1.67$

In diesem Fall ergibt sich ein Testfunktionswert von $z = 15.06$. Bei einem Signifikanzniveau von $\alpha = 0.05$ erhält man den Verwerfungsbereich $B = \langle 7.81; \infty \rangle$, so daß die Nullhypothese verworfen wird.

Für die Distanzmatrix des Anhangs B kann zunächst bei Betrachtung des Falls zufällig fehlender Distanzen (Fall 1), wiederum bei entsprechender Modifikation der Hypothesen und des Testablaufs, die folgende Arbeitstabelle aufgestellt werden:

Objekt i	1	2	3	4	5	6	7	9	10
h_i	0	0	0	0	0	0	0	0	0
$\tilde{h}_i$	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3

$h^{nom} = 1$	$h^{mom} = 9$
$\tilde{h}^{nom} = 1.04$	$\tilde{h}^{mom} = 6.26$

Der Testfunktionswert beträgt 3.90. Die Nullhypothese kann bei einem Signifikanzniveau von $\alpha = 0.05$ und dem daraus resultierenden Verwerfungsbereich $B = \langle 18.31; \infty \rangle$ nicht abgelehnt werden.

Betrachtet man abschließend noch den für die Distanzmatrix im Anhang B beschriebenen Fall 3, dann ergibt sich die folgende Arbeitstabelle:

Objekt i	1	2	3	6	10
h_i	0	0	0	0	0
$\tilde{h}_i$	0.42	0.42	0.42	0.42	0.42

$h^{nom} = 5$	$h^{mom} = 5$
$\tilde{h}^{nom} = 0.53$	$\tilde{h}^{mom} = 7.37$

Als Testfunktionswert erhält man einen Wert von 40.56. Bei einem Signifikanzniveau von $\alpha = 0.05$ und dem daraus resultierenden Verwerfungsbereich $B = \langle 12.59; \infty \rangle$ wird die Nullhypothese verworfen. Ergänzend sei an dieser Stelle angemerkt, daß die maximal mögliche weitere Aufteilung der erwarteten sowie beobachteten Anzahl an Objekten mit mindestens zwei fehlenden Distanzen in die jeweilige Anzahl an Objekten mit genau zwei, genau drei, genau vier und genau fünf fehlenden Distanzen in diesem Fall zu einem Testfunktionswert von 60.30 führt.

Den Ausgangspunkt für eine weitere Testvariante zur Untersuchung der Abhängigkeit der fehlenden Werte vom Fehlen der Daten bei anderen Merkmalen bzw. Objekten stellen die in Abschnitt 3.2.1 angegebenen, auf Basis der Indikatormatrizen berechneten Korrelationsmatrizen dar. Entsprechend können **Tests bezüglich der Korrelationskoeffizienten** zur Anwendung kommen. Dabei kann zum einen die Hypothese überprüft werden, ob die gemäß (3.35) bis (3.37) bestimmten Korrelationskoeffizienten signifikant von Null abweichen. In diesem Fall besteht die Möglichkeit, entweder einzelne Koeffizienten der Korrelationsmatrizen $R^{V,M}$, $R^{V,N}$ bzw. R^{W} zu testen oder jeweils alle Korrelationskoeffizienten gleichzeitig einer Überprüfung zu unterziehen (vgl. z.B. *Hartung, Elpelt, 1992, S. 153-154, 162-164*). Zum anderen kann die Hypothese getestet werden, ob die Korrelationen, die zwischen den Merkmalen bzw. Objekten auf Basis der Indikatormatrix und den gemäß (3.1), (3.2) bzw. (3.15) bestimmten MD-Indikatorvariablen berechnet werden können, homogen sind. Dabei wird also die Gleichheit der Koeffizienten überprüft, die sich jeweils in den Korrelationsmatrizen $R^{V,M+}$, $R^{V,N+}$ bzw. R^{W+} in der gegenüber den Matrizen $R^{V,M}$, $R^{V,N}$ bzw. R^{W} zusätzlich hinzugefügten Zeile bzw. Spalte befinden. Ein entsprechendes Testverfahren ist beispielsweise bei *Hartung und Elpelt (1992, S. 160-161)* beschrieben.

Beispiel:

Für die Datenmatrix des Anhangs A soll lediglich der Fall zufällig fehlender Daten (Fall 1) berücksichtigt werden. Bei einer ausschließlich merkmalsbezogenen Betrachtung stellt damit die in Abschnitt 3.2.1 bereits ermittelte Korrelationsmatrix $R^{V,M}$ den Ausgangspunkt dar. Es soll die Nullhypothese, daß alle Korrelationskoeffizienten gleich Null sind, gegen die Alternativhypothese, daß mindestens ein Korrelations-

koeffizient ungleich Null ist, getestet werden. Dazu ist ein Globaltest durchzuführen, wobei als Teststatistik beispielsweise der Ausdruck

$$W = \left(\tfrac{4}{3}\cdot(m-q)-n+\tfrac{5}{6}\right)\cdot \ln\left(\det R^{V,M}\right)$$

verwendet werden kann (vgl. *Hartung, Elpelt, 1992, S. 163-164*). Für W ergibt sich in diesem Beispiel ein Wert von 13.22. Die Nullhypothese kann bei einem Signifikanzniveau von $\alpha = 0.05$ nicht abgelehnt werden, da der Testfunktionswert kleiner als das 0.95-Fraktil der $\chi^2(15)$-Verteilung mit 25.00 ist. Die Anzahl der Freiheitsgrade df ergibt sich dabei gemäß

$$df = \frac{(m-q)\cdot(m-q-1)}{2}.$$

Damit kann die bereits im entsprechenden Beispiel des Abschnitts 3.2.1 geäußerte Vermutung eines unsystematischen Fehlens der Daten statistisch untermauert werden.

Mit dem dritten Testansatz, der grundsätzlich nur bei Vorliegen einer unvollständigen Datenmatrix relevant ist, soll die Abhängigkeit der fehlenden Werte von den vorhandenen Ausprägungen bei anderen Merkmalen bzw. Objekten untersucht werden. Im Fall einer merkmalsweisen Betrachtung werden dazu für jedes Merkmal $k \in M$ mit fehlenden Daten, d.h. $v_{\bullet k}^{ind} = 0$, gemäß den Werten der zugehörigen MD-Indikatorvariablen die vorhandenen Ausprägungen von jedem Merkmal $l \in M, l \neq k$ in zwei Gruppen unterteilt. Diese Gruppen sind dann jeweils auf signifikante Unterschiede zu untersuchen (vgl. z.B. *Little, 1988, S. 1198*). Dazu eignen sich abhängig vom Skalenniveau des Merkmals l **Tests auf Lokationsunterschiede bzw. Unabhängigkeit** (vgl. *Möntmann et al., 1983, S. 90-95*). Bei den Tests auf Lokationsunterschiede wird ein Vergleich der Lageparameter der beiden Gruppen durchgeführt. Bei kardinalem Skalenniveau des Merkmals l können, je nach Verteilungsannahmen über die Ausprägungen, der **Zweistichproben-Gaußtest**, der **Zweistichproben-t-Test** sowie der **approximative Zweistichproben-Gaußtest** zur Anwendung kommen (vgl. z.B. *Bamberg, Baur, 1993, S. 192-194*). Im Fall eines ordinal skalierten Merkmals l ist der **Zweistichproben-Vorzeichentest** geeignet (vgl. z.B. *Hartung, 1989, S. 540-541*). Bei den Tests auf Unabhängigkeit zwischen der zum Merkmal k gehörenden MD-Indikatorvariable und einem Merkmal $l \in M, l \neq k$ wird lediglich nominales Skalenniveau für das Merkmal l vorausgesetzt. Als Beispiel sei hier der **χ^2-Test** genannt (vgl. z.B. *Hartung, 1989, S. 413-414*).

Für jedes Merkmal $k \in M$ mit fehlenden Daten können $(m-1)$ und für die unvollständige Datenmatrix damit insgesamt

$$(m-q)\cdot(m-1) \quad \text{mit} \quad q = \sum_{h=1}^{m} v_{\bullet h}^{ind} \tag{3.50}$$

Tests auf Lokationsunterschiede bzw. Unabhängigkeit durchgeführt werden. Eine vollständige Untersuchung der Abhängigkeit der MD von den Ausprägungen bei anderen

Merkmalen ist mit dem vorgestellten Testansatz also sehr aufwendig und in praktischen Fragestellungen in der Regel nicht zweckmäßig. *Frane (1978, S. 27)* schlägt daher vor, lediglich für das Merkmal mit den meisten fehlenden Daten gemäß der entsprechenden MD-Indikatorvariablen eine Aufteilung der vorhandenen Ausprägungen der anderen Merkmale durchzuführen und diese Gruppen auf signifikante Unterschiede zu untersuchen.

Falls im Rahmen der Tests auf Lokationsunterschiede bzw. Unabhängigkeit signifikante Unterschiede zwischen den Gruppen der Merkmale $l \in M$, $l \neq k$, die jeweils durch die zum Merkmal $k \in M$ gehörende MD-Indikatorvariable gebildet werden, auftreten, ist eine Annahme der Eigenschaft OAR für die Daten nicht tragbar. Der umgekehrte Fall stellt jedoch wiederum nur eine notwendige Bedingung zur Akzeptierung der Eigenschaft OAR und damit auch des unsystematischen Fehlens der Daten dar.

Analog zu der oben dargestellten merkmalsbezogenen Betrachtung kann eine objektbezogene Analyse durchgeführt werden. Dies ist allerdings nur dann möglich, wenn in der Datenmatrix ein einheitliches Skalenniveau mit vergleichbaren Ausprägungsmengen vorliegt.

Beispiel:

In der Datenmatrix des Anhangs A weisen für den Fall zufällig fehlender Daten (Fall 1) die Merkmale Deskriptive Statistik und Spezialgebiete mit jeweils drei MD den höchsten Anteil fehlender Daten auf. Im folgenden soll lediglich für das Merkmal Deskriptive Statistik untersucht werden, inwieweit die fehlenden Daten bei diesem Merkmal von den Daten der Merkmale Testverfahren, Multivariate Verfahren, Businessgrafiken und Statistische Grafiken abhängig sind. Durch die zum Merkmal Deskriptive Statistik gehörende MD-Indikatorvariable werden die vorhandenen Daten der Merkmale Testverfahren, Multivariate Verfahren, Businessgrafiken und Statistische Grafiken jeweils in zwei Gruppen eingeteilt. Unter der Annahme, daß die Ausprägungen der auf diese Art gebildeten Gruppen jeweils einer Normalverteilung genügen und die Varianzen innerhalb der beiden Gruppen jeweils identisch, aber unbekannt sind, stellt der Zweistichproben-t-Test ein geeignetes Testverfahren dar. Bei einem Signifikanzniveau von $\alpha = 0.05$ erhält man die folgenden Ergebnisse der vier durchzuführenden Tests, wobei jeweils die Gleichheit bzw. Ungleichheit der Erwartungswerte innerhalb der beiden Gruppen überprüft wird:

Merkmal	Testverfahren	Multivariate Verfahren	Businessgrafiken	Statistische Grafiken
Testfunktionswert	−0.53	−1.18	−0.03	−1.19
Verwerfungsbereich	$\langle -\infty; -2.16\rangle \cup \langle 2.16; \infty\rangle$	$\langle -\infty; -2.16\rangle \cup \langle 2.16; \infty\rangle$	$\langle -\infty; -2.16\rangle \cup \langle 2.16; \infty\rangle$	$\langle -\infty; -2.20\rangle \cup \langle 2.20; \infty\rangle$

Ein signifikanter Unterschied zwischen den Gruppen, die jeweils durch die zum Merkmal Deskriptive Statistik gehörende MD-Indikatorvariable gebildet werden, ist nicht feststellbar. Damit kann auf eine Ab-

hängigkeit der MD beim Merkmal Deskriptive Statistik von den Daten der Merkmale Testverfahren, Multivariate Verfahren, Businessgrafiken und Statistische Grafiken nicht geschlossen werden.

Ergänzend wird im folgenden der im Anhang A beschriebenen Fall 4, bei dem das Fehlen der Daten beim Merkmal Preisniveau von den Ausprägungen beim Merkmal Deskriptive Statistik abhängt, betrachtet. Mit der MD-Indikatorvariable des Merkmals Preisniveau werden die Ausprägungen des Merkmals Deskriptive Statistik in zwei Gruppen eingeteilt, wobei wiederum eine Normalverteilung mit identischer, aber unbekannter Varianz für die Ausprägungen beider Gruppen unterstellt wird. Die Anwendung des Zweistichproben-t-Tests führt zu einem Testfunktionswert von 7.70. Bei einem Signifikanzniveau von $\alpha = 0.05$ und einem daraus resultierenden Verwerfungsbereich von $\langle -\infty;-2.16\rangle \cup \langle 2.16;\infty\rangle$ liegt also ein signifikanter Unterschied zwischen den Erwartungswerten der beiden Gruppen vor. Daraus resultiert, daß in diesem Fall die Annahme OAR für die Daten nicht gerechtfertigt ist.

Ein weiteres Testverfahren des dritten Testansatzes stellt der **Test nach *Little*** dar (vgl. *Little, 1988*). Dieser Test hat den Vorteil, daß im Vergleich zu den bisher beschriebenen Tests auf Lokationsunterschiede bzw. Unabhängigkeit lediglich eine einzige Teststatistik für die gesamte Datenmatrix und damit für eine vollständige Untersuchung der Abhängigkeit der MD von den Ausprägungen bei anderen Merkmalen bestimmt werden muß. Der Grundgedanke des Tests nach *Little* besteht darin, die jeweils berechenbaren Merkmalsmittelwerte, die sich auf Basis aller Objekte mit einem identischen MD-Muster ergeben, auf signifikante Abweichungen von den Gesamtmittelwerten dieser Merkmale im Fall vollständiger Daten zu untersuchen. Voraussetzung dabei ist, daß in der Datenmatrix lediglich kardinal skalierte Merkmale, deren Ausprägungen einer Normalverteilung genügen, vorliegen.

Sei H die Anzahl der verschiedenen MD-Objektmuster, n_h die Anzahl der Objekte, die das MD-Objektmuster h ($h = 1,...,H$) aufweisen, m_h die Anzahl aus den der Datenmatrix insgesamt zugrundeliegenden m Merkmalen, deren Ausprägungen bei dem MD-Objektmuster h vorhanden sind, $\bar{a}^{h,obs} \in \mathbb{R}^{m_h \times 1}$ der Mittelwertvektor der zum MD-Objektmuster h gehörenden Objektvektoren a^i, dessen Werte lediglich für die m_h vollständigen Merkmale bestimmt werden, sowie $\bar{a}^h \in \mathbb{R}^{m_h \times 1}$ bzw. $S^h \in \mathbb{R}^{m_h \times m_h}$ der Teil des Mittelwertvektors bzw. der Kovarianzmatrix für den Fall einer vollständigen Datenmatrix, der die m_h Merkmale, deren Ausprägungen beim MD-Objektmuster h vorhanden sind, betrifft, dann ergibt sich die bei *Little (1988, S. 1200)* mit d^2 bezeichnete Teststatistik gemäß der Formel

$$d^2 = \sum_{h=1}^{H} n_h \left(\bar{a}^{h,obs} - \bar{a}^h\right)^T \cdot \left(S^h\right)^{-1} \cdot \left(\bar{a}^{h,obs} - \bar{a}^h\right). \tag{3.51}$$

Diese Teststatistik, die in Anlehnung an die Hotellingsche-T^2-Statistik (vgl. z.B. *Hartung, 1989, S. 227*) konstruiert wird, ist unter der Annahme einer Normalverteilung für

die Ausprägungen der zugrundeliegenden m Merkmale näherungsweise χ^2-verteilt, wobei sich die Anzahl der Freiheitsgrade df gemäß

$$df = -m + \sum_{h=1}^{H} m_h \tag{3.52}$$

ergibt (vgl. *Little, 1988, S. 1200*). Es stellt sich die Problematik, daß der Mittelwertvektor sowie die Kovarianzmatrix für den Fall vollständiger Daten nicht bekannt sind. Gemäß *Little (1988, S. 1200)* müssen also entsprechende Schätzwerte herangezogen werden, zu deren Bestimmung allerdings bereits Verfahren zur Berücksichtigung der fehlenden Daten anzuwenden sind. Da der Testansatz jedoch implizit eine Überprüfung der OAR-Annahme für die Daten enthält, können lediglich Verfahren zur Behandlung der fehlenden Daten verwendet werden, die auf der MAR-Annahme für die Daten und nicht auf der restriktiveren MCAR-Annahme, in der die OAR-Annahme enthalten ist, basieren. Die damit in Frage kommenden Verfahren zur Bestimmung der Schätzwerte für den Mittelwertvektor und die Kovarianzmatrix des gesamten Datenmaterials werden in Kapitel 4 dieser Arbeit noch ausführlich dargestellt.

Zusammenfassend ergibt sich damit die folgende Grundkonzeption des Tests nach *Little*: Ausgehend von der Annahme MAR für die Daten, gegen deren Akzeptierung keine Einwände bestehen, werden geeignete Verfahren zur Schätzung des Mittelwertvektors und der Kovarianzmatrix des gesamten Datenmaterials angewandt. Sind die Abweichungen zwischen den jeweils berechenbaren Merkmalsmittelwerten, die sich auf Basis aller Objekte mit einem identischen MD-Muster ergeben, und den geschätzten Gesamtmittelwerten signifikant, dann ist die OAR-Annahme und damit auch die MCAR-Annahme für Daten nicht tragbar. Der umgekehrte Fall, d.h. die Abweichungen sind nicht signifikant, stellt hingegen lediglich eine notwendige Bedingung zur Akzeptierung der zusätzlichen Annahme OAR und damit insgesamt der Annahme MCAR für die Daten dar.

Der oben dargestellte Testansatz kann auch dahingehend modifiziert werden, daß eine Analyse der Abhängigkeit der fehlenden Daten von den vorhandenen Ausprägungen bei anderen Objekten möglich ist. In diesem Fall müssen in der Datenmatrix jedoch wiederum ein einheitliches Skalenniveau sowie vergleichbare Ausprägungsmengen vorliegen.

Im Hinblick auf die praktische Anwendung kann für den Test nach *Little* festgehalten werden, daß die Untersuchung der Eigenschaft OAR mittels einer einzigen Teststatistik für die gesamte Datenmatrix sehr vorteilhaft ist. Die Nachteile der ausschließlich möglichen Verwendung kardinaler Merkmale sowie der notwendigen Schätzung des Mittelwertvektors und der Kovarianzmatrix schränken jedoch die Anwendung dieses Tests stark ein.

Beispiel:

Für die sechs kardinalen Merkmale der Datenmatrix des Anhangs A soll der Test nach *Little* durchgeführt werden. Dabei wird für die Ausprägungen dieser Merkmale eine Normalverteilung unterstellt. Des weiteren werden zur Vereinfachung der Mittelwertvektor und die Kovarianzmatrix der vollständigen Datenmatrix verwendet, die sich wie folgt ergeben:

$$\bar{a} = \begin{pmatrix} 77.93 \\ 72.53 \\ 35.45 \\ 44.97 \\ 54.07 \\ 37.82 \end{pmatrix}, \quad S = \begin{pmatrix} 160.29 & 17.47 & 183.81 & 160.99 & 176.25 & 25.38 \\ 17.47 & 127.41 & -32.92 & 24.25 & 42.18 & -82.47 \\ 183.81 & -32.92 & 470.71 & 437.99 & 545.49 & 296.04 \\ 160.99 & 24.25 & 437.99 & 703.92 & 812.48 & 417.43 \\ 176.25 & 42.18 & 545.49 & 812.48 & 1716.35 & 750.46 \\ 25.38 & -82.47 & 296.04 & 417.43 & 750.46 & 636.77 \end{pmatrix}.$$

Im betrachteten Teil der Datenmatrix liegen insgesamt $H = 5$ verschiedene MD-Objektmuster vor. Im folgenden sind die einzelnen MD-Objektmuster h (h = 1,...,5) einschließlich der jeweiligen Werte n_h und m_h angegeben, wobei im MD-Objektmuster eine vorhandene Ausprägung mit 1 und eine fehlende Ausprägung mit 0 angedeutet wird:

MD-Objektmuster						h	n_h	m_h
1	1	1	1	1	1	1	9	6
0	1	1	1	1	1	2	1	5
1	1	1	1	1	0	3	2	5
0	1	1	0	1	1	4	2	4
1	1	1	0	1	1	5	1	5

Im Hinblick auf die Berechnung der Teststatistik d^2 sind im folgenden beispielhaft für das MD-Objektmuster $h = 4$ die Vektoren $\bar{a}^{4,obs}$ und $\bar{a}^4$ sowie die Matrix S^4 angegeben:

$$\bar{a}^{4,obs} = \begin{pmatrix} 73.00 \\ 25.71 \\ 70.00 \\ 31.63 \end{pmatrix}, \quad \bar{a}^4 = \begin{pmatrix} 72.53 \\ 35.45 \\ 54.07 \\ 37.82 \end{pmatrix}, \quad S^4 = \begin{pmatrix} 127.41 & -32.92 & 42.18 & -82.47 \\ -32.92 & 470.71 & 545.49 & 296.04 \\ 42.18 & 545.49 & 1716.35 & 750.46 \\ -82.47 & 296.04 & 750.46 & 636.77 \end{pmatrix}.$$

Als Testfunktionswert erhält man $d^2 = 3.64 + 4.86 + 2.02 + 3.13 + 4.94 = 18.59$. Bei einem Signifikanzniveau von $\alpha = 0.05$ ergibt sich der Wert des zugehörigen 0.95-Fraktils der $\chi^2(19)$-Verteilung mit 30.14. Damit spricht nichts gegen die Annahme OAR für den verwendeten Teil der Daten.

Für den im Anhang A beschriebenen Fall 4, bei dem die Abhängigkeit der fehlenden Daten von den vorhandenen Ausprägungen eines anderen Merkmals berücksichtigt wird, kann der Test nach *Little* nicht angewandt werden. Im betrachteten Bereich der Datenmatrix liegt nicht ausschließlich ein kardinales Skalenniveau vor.

3.4 Konsequenzen für die Behandlung der fehlenden Daten

Die Ergebnisse einer Auswertung unvollständiger Daten- und Distanzmatrizen werden im allgemeinen fehlerbehaftet sein, wenn der den Daten zugrundeliegende Ausfallmechanismus systematisch und unbekannt ist (vgl. z.B. *Greenless et al., 1982, S. 252*). So können bei achtloser Verwendung der vorliegenden Daten ganze Subgruppen aus der Analyse ausgeschlossen werden und somit falsche Schätzwerte entstehen. Zu denken ist hier beispielsweise an eine Erhebung bezüglich des Einkommens, bei der die Daten von Personen mit hohem Einkommen aufgrund einer Antwortverweigerung nicht erfaßt werden.

Eine adäquate Behandlung der fehlenden Daten und somit korrekte Analyseergebnisse im Fall eines systematischen Ausfallmechanismus sind nur bei genauer Kenntnis des zu den fehlenden Daten führenden Mechanismus möglich, da in diesem Fall eine Modellierung des Ausfallmechanismus denkbar ist (vgl. *Schnell, 1986, S. 10*). Diese Vorgehensweise ist jedoch generell als schwierig anzusehen und in praktischen Anwendungen oft nicht durchführbar, zumal bislang nur wenige Ansätze zur Behandlung einer unvollständigen Daten- oder Distanzmatrix bei einem systematischen, aber zumindest bekannten Ausfallmechanismus existieren. Erheblich einfacher ist hingegen der Fall, daß der den Daten zugrundeliegende Ausfallmechanismus als vernachlässigbar bzw. ignorierbar angesehen werden kann. Grundsätzlich ist dies dann möglich, wenn die beobachteten Werte einer Daten- oder Distanzmatrix eine zufällige Stichprobe der Gesamtstichprobe darstellen und die nicht vorhandenen Werte zufällig fehlen (vgl. *Toutenburg, 1992, S. 200*). Der Großteil der in der Literatur beschriebenen Ansätze und Verfahren zur Behandlung einer unvollständigen Daten- oder Distanzmatrix geht von einem unsystematischen und damit vernachlässigbaren Ausfallmechanismus aus. Abhängig von den in Abschnitt 2.2 dargestellten Formen eines unsystematischen Ausfallmechanismus kann jedoch eine genaue Unterscheidung getroffen werden, welche Verfahren im einzelnen angewandt werden können. Diese Überlegungen gehen auf eine Arbeit von *Rubin (1976)* zurück und sollen im folgenden kurz vorgestellt werden.

Den Ausgangspunkt der Betrachtung stellen die Bedingungen MAR und MCAR für die Daten dar. Nach *Rubin (1976, S. 585)* sind die Folgerungen aus den vorliegenden Daten, die als Stichprobe aus dem vollständigen Datenmaterial angesehen werden kann, korrekt, wenn die Daten der Eigenschaft MCAR genügen. In diesem Fall können zur Behandlung der fehlenden Daten Verfahren angewandt werden, die auf den vorhandenen Daten und damit auf der vorliegenden Stichprobe des zu erhebenden Datenmaterials basieren. Falls die Daten lediglich der im Vergleich zu MCAR schwächeren Eigenschaft MAR genügen, sind unmittelbare Folgerungen aus der vorliegenden Stichprobe zwar nicht korrekt, jedoch führen Folgerungen unter Verwendung der Maximum-Likelihood-

sowie der Bayes-Theorie zu einem richtigen Ergebnis (vgl. *Rubin, 1976, S. 586-587*).[4] Damit können bei Daten, für die lediglich die Annahme MAR und nicht die Annahme OAR gerechtfertigt ist, Verfahren zur Behandlung der fehlenden Werte angewandt werden, die auf der Likelihood- bzw. der Bayes-Theorie basieren.

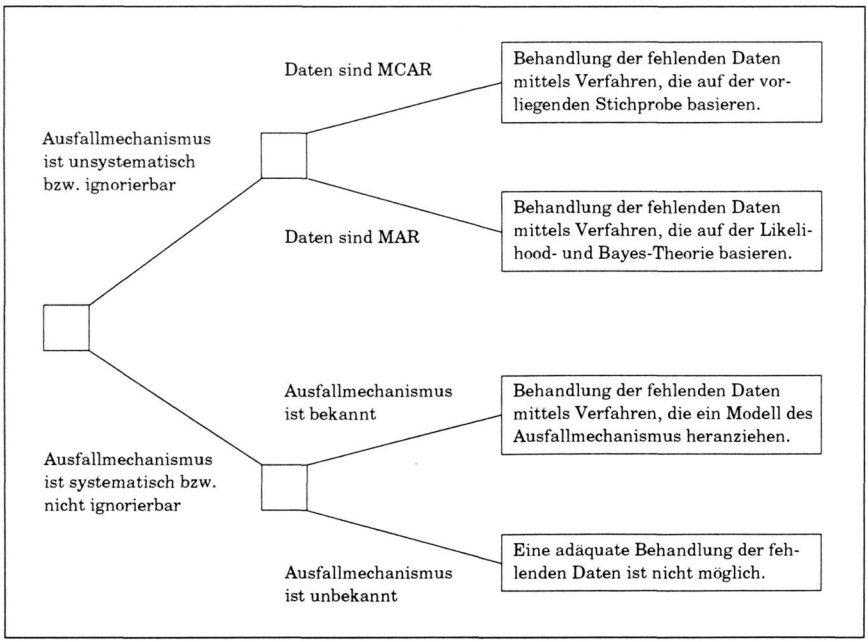

Abbildung 3.20: Konsequenzen des Ausfallmechanismus für die Datenauswertung

Zusammenfassend sind in der Abbildung 3.20 noch einmal die Konsequenzen für die Behandlung der fehlenden Daten in Abhängigkeit des zugrundeliegenden Ausfallmechanismus in Form eines Entscheidungsbaums dargestellt. Mittels der in diesem Kapitel vorgestellten Ansätze zur Strukturanalyse einer unvollständigen Daten- oder Distanzmatrix sollte also zunächst festgelegt werden, ob die Annahme eines unsystematischen Ausfallmechanismus für die Daten grundsätzlich gerechtfertigt ist bzw. die Daten nicht gegen diese Annahme sprechen. Ist dies der Fall, dann ist in Abhängigkeit von den Annahmen MCAR und MAR eine Behandlung der fehlenden Daten mittels der jeweils ge-

[4] Von *Rubin (1976, S. 585-586)* wird zusätzlich die Bedingung gefordert, daß die Parameter der Verteilung für die zur unvollständigen Daten- bzw. Distanzmatrix gehörenden Indikatormatrix von den aus dem unvollständigen Datenmaterial zu schätzenden Parameter verschieden sein müssen. Diese Bedingung wird im folgenden nicht weiter berücksichtigt und stets als gegeben vorausgesetzt, da sie im Rahmen dieser Arbeit nicht von Bedeutung ist.

eigneten Verfahren möglich. Liegt hingegen ein nicht ignorierbarer Ausfallmechanismus vor, dann können die vorgestellten Ansätze zur Strukturanalyse darüber hinaus dazu beitragen, die Art des Ausfallmechanismus genauer zu spezifizieren und aufzudecken. In diesem Fall kann unter Umständen eine Modellierung des Mechanismus, der die fehlenden Daten erzeugt, durchgeführt werden. Auf Basis dieses Modells ist dann eine Behandlung der fehlenden Daten unter Anwendung der entsprechenden Verfahren möglich. Bleibt jedoch die Art des Ausfallmechanismus unbekannt, dann ist eine adäquate Behandlung der fehlenden Werte im Rahmen der Datenauswertung nicht durchführbar. Abhängig von unterschiedlichen Annahmen über den zugrundeliegenden Ausfallmechanismus kann lediglich die Spannweite der möglichen Ergebnisse aufgezeigt werden.

4 Verfahren zur Behandlung fehlender Daten

In der Literatur existieren mittlerweile eine Reihe unterschiedlicher Möglichkeiten, das Problem fehlender Daten zu behandeln. In den bislang veröffentlichten Arbeiten werden die sogenannten **MD-Verfahren** (Missing-Data-Verfahren) jedoch meist unstrukturiert und ohne Berücksichtigung eines übergeordneten Untersuchungsziels dargestellt. Darüber hinaus beschränken sich diese Veröffentlichungen häufig auf Verfahren, die eine Normalverteilung der Variablen sowie ein zufälliges Fehlen der Daten voraussetzen.

In diesem Kapitels sollen die aus der Literatur bekannten sowie die daraus ableitbaren bzw. darüber hinaus denkbaren MD-Verfahren vorgestellt werden, die für eine Anwendung im Rahmen der Datenanalyse geeignet sind. Neben der grundsätzlichen Verträglichkeit mit der Zielsetzung einer datenanalytischen Untersuchung hängt die Eignung eines MD-Verfahrens vor allem von den jeweils zugrundeliegenden Annahmen ab. Dazu zählen Annahmen über den Ausfallmechanismus, die Verteilung der Merkmalsausprägungen sowie das zulässige Skalenniveau der Merkmale. Für die Datenanalyse ergeben sich lediglich Einschränkungen bezüglich spezieller Verteilungsannahmen. Bei der Darstellung der Verfahren ist dieser Aspekt daher besonders zu berücksichtigen. Darüber hinaus sollen mögliche Lösungsansätze gerade im Hinblick auf systematische Ausfallmechanismen sowie beliebige Skalenniveaus der Merkmale erarbeitet werden.

Ausgangspunkt der nachfolgenden Betrachtungen ist grundsätzlich eine Datenmatrix der Form (1.1), da auch die in der Literatur bekannten MD-Verfahren fast ausnahmslos auf einer derartigen Matrix basieren. Denkbare Modifikationen sowie eigene Ansätze im Hinblick auf eine unmittelbar erhobene Distanzmatrix der Form (1.3) sollen daher jeweils explizit erwähnt werden. In Anlehnung an *Schwab (1991, S. 4), Beale und Little (1975, S. 130)* sowie *Frane (1976, S. 409)* können fünf Strategien zur Behandlung fehlender Daten unterschieden werden, so daß eine Einteilung der MD-Verfahren in die folgenden Kategorien vorgenommen wird:

- **Eliminierungsverfahren**
- **Imputationsverfahren**
- **Parameterschätzverfahren**
- **Multivariate Analyseverfahren**
- **Sensitivitätsbetrachtungen**

Jeder dieser fünf Verfahrenskategorien ist in diesem Kapitel ein eigener Abschnitt gewidmet. Neben einer ausführlichen Darstellung der einzelnen Verfahren soll vor allem

eine Bewertung anhand der jeweiligen Voraussetzungen und Eigenschaften sowie der daraus resultierenden Konsequenzen für die Eignung im Rahmen einer datenanalytischen Untersuchung erfolgen.

Das Entfernen von Objekten bzw. Merkmalen mit fehlenden Werten aus der Untersuchung wird in **Abschnitt 4.1** als einfachste Methode zur Lösung des Problems fehlender Daten dargestellt. Die Auswertung kann dann auf Basis des reduzierten, aber vollständigen Datenmaterials durchgeführt werden. In **Abschnitt 4.2** werden verschiedene Techniken der Imputation fehlender Werte beschrieben. Bei diesen Verfahren wird das Datenmaterial durch geeignete Schätzwerte für die fehlenden Ausprägungen bzw. Distanzen vervollständigt. Entsprechend der Zielsetzung der Analyse kann dann ein herkömmliches Auswertungsverfahren, das eine vollständige Daten- bzw. Distanzmatrix voraussetzt, zur Anwendung kommen. Der **Abschnitt 4.3** widmet sich ausgewählten Schätzverfahren, mit denen Parameter, wie beispielsweise Mittelwerte oder Varianzen bzw. Kovarianzen, aus dem unvollständigen Datenmaterial bestimmt werden. Diese Ergebnisse können dann in einigen multivariaten Verfahren, wie z.B. der Faktoren- oder Diskriminanzanalyse, herangezogen werden. In **Abschnitt 4.4** werden multivariate Analyseverfahren vorgestellt, die unmittelbar auf der unvollständigen Daten- bzw. Distanzmatrix basieren und ausschließlich die vorliegenden Daten verwenden. Diese Verfahren ergeben sich im allgemeinen durch geringfügige Modifikationen der entsprechenden, auf vollständigen Daten basierenden Analysemethoden. Der **Abschnitt 4.5** behandelt schließlich die Möglichkeiten einer Sensitivitätsanalyse auf Basis der unvollständigen Daten- oder Distanzmatrix. Bei diesen Ansätzen wird die Sensitivität von Imputationswerten, Parameterschätzungen oder multivariaten Analyseergebnissen gegenüber dem verwendeten MD-Verfahren oder dem zugrundegelegten Modell des Ausfallmechanismus untersucht.

Da sich einige MD-Verfahren, die beispielsweise neben der Schätzung der Kovarianzmatrix auch Imputationswerte liefern, nicht eindeutig in eine der fünf Kategorien einteilen lassen, erfolgt die Zuordnung nach der grundsätzlichen Intention dieser Verfahren. Bei der jeweiligen Darstellung wird jedoch auf die Möglichkeit der Durchführung weiterer Strategien hingewiesen.

Der an sich naheliegende Ansatz, daß die Datenauswertung unmittelbar auf Basis einer Daten- bzw. Distanzmatrix erfolgt, in der alle nicht vorhandenen Werte als fehlend codiert sind, wird im folgenden nicht weiter betrachtet. Dies ist damit zu begründen, daß ein entsprechender Codierungswert lediglich bei den nominalen Merkmalen einer Datenmatrix als eigenständige Ausprägung betrachtet werden kann und im Fall einer Distanzmatrix oder bei den kardinalen und ordinalen Merkmalen einer Datenmatrix nicht mit den vorhandenen Daten verträglich ist.

4.1 Eliminierungsverfahren

Unter dem Begriff Eliminierungsverfahren sind Methoden zur Behandlung unvollständiger Daten- oder Distanzmatrizen zusammengefaßt, die Objekte bzw. Merkmale mit fehlenden Werten aus der Untersuchung ausschließen. Diese Verfahren sind nur unter der sehr restriktiven Annahme, daß das Fehlen einzelner Werte in keiner Beziehung zu den Variablen der Untersuchung steht, d.h. die Daten der Bedingung MCAR genügen, uneingeschränkt anwendbar (vgl. *Little, Rubin, 1987, S. 39-40*). Weitere Voraussetzungen werden an das Datenmaterial jedoch nicht gestellt, so daß die Verteilung der Merkmalsausprägungen sowie das Skalenniveau der Merkmale beliebig sein können. Bei Vorliegen einer unmittelbar erhobenen Distanzmatrix ist lediglich die im folgenden zunächst dargestellte Eliminierung von Objekten in Betracht zu ziehen.

4.1.1 Objekteliminierung

Falls die Struktur der Merkmale untersucht werden soll, stellt eine Objekteliminierung, also der Ausschluß von Objekten mit fehlenden Daten aus der Untersuchung, eine durchaus zweckmäßige Lösung dar (vgl. *Frane, 1976, S. 409*). Im Rahmen einer Analyse der Ähnlichkeitsstruktur der Objekte ist eine Objekteliminierung jedoch nur dann geeignet, wenn die Anzahl der zu eliminierenden Objekte verhältnismäßig gering ist. Bei bestimmten Konstellationen in der Daten- oder Distanzmatrix, wie beispielsweise lediglich einem Objekt mit fehlenden Merkmalsausprägungen, kann eine Objekteliminierung somit auch im Fall einer taxonomischen Aufgabenstellung eine sinnvolle Vorgehensweise darstellen.

4.1.1.1 Analyse der vollständigen Objekte

Im Rahmen einer Auswertung der vollständig erhobenen Objekte (**complete-case analysis**) werden nur die Objekte in einer Analyse verwendet, deren Merkmalsausprägungen bezüglich aller Merkmale vorliegen. Wird die Datenmatrix in der Form

$$A = (a_{ik})_{n,m} = \begin{pmatrix} A_{obs} \\ A_{mis} \end{pmatrix} = \begin{pmatrix} (a_{ik})_{p,m} \\ (a_{ik})_{n-p,m} \end{pmatrix} \tag{4.1}$$

geschrieben, wobei A_{obs} die p vollständigen und A_{mis} die $(n - p)$ unvollständigen Objektvektoren enthält, dann wird im Rahmen einer complete-case analysis nur die Matrix A_{obs} herangezogen, die im Vergleich zur Datenmatrix A die niedrigere Dimension $(p \times m)$ besitzt.

Beispiel:

Für die Datenmatrix des Anhangs A (Fall 1) ergibt sich nach diesem Verfahren eine auswertbare Matrix A_{obs} mit der Dimension (7 × 9), wobei die Objekte CRUNCH, MICROSTAT II, SAS, SPSS, STATA, STATGRAPHICS, STATPAC GOLD und SYSTAT aus der weiteren Untersuchung eliminiert werden. Zur Durchführung einer Analyse der Ähnlichkeitsstruktur der Objekte stellt die Objekteliminierung in diesem Fall jedoch keine befriedigende Lösung dar.

Falls die Datenmatrix einen geringen Prozentsatz fehlender Werte aufweist (< 5%), wird die Durchführung einer Objekteliminierung von *Schwab (1991, S. 4-5)* aus Kosten- und Zeitersparnisgründen als akzeptabel bezeichnet. Bei höheren Prozentsätzen fehlender Werte in einer Datenmatrix kann dieses Verfahren jedoch zu einem erheblichen Informationsverlust führen, wie folgendes Beispiel, in dem K die Anzahl der Variablen darstellt, verdeutlicht:

> *"..., if $K = 20$ and each variable is missing or observed independently according to a Bernoulli process with a 10% chance of missingness, then the expected proportion of complete cases is $0{,}9^{20} = 0{,}12$. That is, only about $12/0{,}9 = 13$ percent of the observed data values will be retained." [Little, Rubin, 1987, S. 40]*

Falls eine unvollständige Distanzmatrix direkt erhoben wurde, kann die complete-case analysis sinngemäß angewandt werden. Dabei wird zunächst das Objekt eliminiert, das die höchste Anzahl fehlender Distanzen zu anderen Objekten aufweist, dann das mit der zweithöchsten Anzahl usw., bis schließlich eine vollständige Distanzmatrix entsprechend niedrigerer Dimension resultiert. Diese Vorgehensweise führt jedoch nicht notwendig zu einem eindeutigen Ergebnis.

Beispiel:

Für die Distanzmatrix des Anhangs B (Fall 1) existieren mehrere Eliminierungsmöglichkeiten, da neun Objekte mit einem Wert von 2 die höchste Anzahl fehlender Distanzen zu anderen Objekten aufweisen. So können beispielsweise der Reihe nach die Objekte Audi 90, BMW 3er, Ford Sierra, Toyota Carina und Opel Vectra aus der Untersuchung ausgeschlossen werden. Entsprechend ergibt sich dann eine vollständige Distanzmatrix mit fünf Objekten für die weitere Analyse. Alle anderen, gemäß der obigen Heuristik in Frage kommenden Eliminierungsmöglichkeiten führen in diesem Beispiel ebenfalls zu einer vollständigen Distanzmatrix der Dimension (5 × 5).

Die oben angegebene Heuristik zur Bestimmung der zu eliminierenden Objekte im Rahmen einer complete-case analysis einer unvollständigen Distanzmatrix stellt eine sehr einfache Lösungsmethode dar. Verfolgt man jedoch das Ziel, durch die Objekteliminierung eine möglichst geringe Anzahl vorhandener paarweiser Distanzen aus der weiteren Analyse zu entfernen, d.h. einen minimalen Informationsverlust zu erreichen, dann muß ein anderer Lösungsansatz gefunden werden, da die Heuristik nicht unbe-

dingt zu einer Optimallösung gemäß dieser Zielsetzung führt. Dazu wird das folgende Optimierungsproblem formuliert:

$$\text{Informationsverlust} \rightarrow \min \quad \Leftrightarrow \quad \sum_{i=1}^{n} o_i \rightarrow \max \tag{4.2}$$

$$\text{mit} \begin{pmatrix} o_1 \\ \vdots \\ o_n \end{pmatrix} (o_1, \ldots, o_n) \leq \begin{pmatrix} w_{11} & \ldots & w_{1n} \\ \vdots & & \vdots \\ w_{n1} & \ldots & w_{nn} \end{pmatrix} \quad \Leftrightarrow \quad o \cdot o^T \leq W,$$

$$o_i = \begin{cases} 1 & \text{falls Objekt } i \text{ verwendet wird} \\ 0 & \text{sonst,} \end{cases},$$

$$w_{ij} = \begin{cases} 1 & \text{falls Distanz } d_{ij} \text{ vorhanden} \\ 0 & \text{sonst.} \end{cases}.$$

Dieses Optimierungsproblem kann beispielsweise mit Hilfe des allgemeinen **Branch and Bound Prinzips** (vgl. z.B. *Neumann, 1975, S. 333-339*) gelöst werden. Die sich ergebende Lösung muß jedoch nicht unbedingt eindeutig sein, d.h. es können auch mehrere, unterschiedliche Optimallösungen vorliegen.

Beispiel:

Für die Distanzmatrix des Anhangs B (Fall 1) ist die Verwendung von mehr als fünf Objekten im Rahmen einer complete-case analysis nicht möglich. Die im vorherigen Beispiel mit der Heuristik ermittelten Lösungen sind somit im Hinblick auf einen minimalen Informationsverlust bereits optimal und stellen Lösungen des Optimierungsproblems (4.2) dar.

4.1.1.2 Analyse der verfügbaren Objekte

Eine zum vorherigen Abschnitt ähnliche Strategie ist das Verwenden aller jeweils verfügbaren Objekte in einer Analyse (**available-case analysis**). Bei der Berechnung univariater Statistiken werden alle bei den einzelnen Merkmalen jeweils vorhandenen Ausprägungen verwendet. Damit geht keine Information verloren, jedoch variiert der Stichprobenumfang im allgemeinen von Merkmal zu Merkmal (vgl. *Little, Rubin, 1987, S. 41-42*). Zur Berechnung von Korrelationskoeffizienten oder Kovarianzen werden nach diesem Ansatz jeweils nur die Objekte herangezogen, deren Ausprägungen für die beiden betrachteten Merkmale vorhandenen sind. In diesem Fall handelt es sich um die sogenannte **pairwise available-case analysis**. Bei der gleichzeitigen Verwendung aller Merkmale in einer Analyse entspricht die available-case analysis der complete-case analysis. In Anlehnung an *Little, Rubin (1987, S. 42-43), Schnell (1986, S. 84-87) sowie Engelman (1982, S. 114-116)* sollen am Beispiel der Berechnung einer Korrelationsmatrix mögliche Modifikationen dieses Ansatzes dargestellt werden.

Sei $N_{kl} = \{i: v_{ik} = 1 \wedge v_{il} = 1\}$, $N_k = \{i: v_{ik} = 1\}$ und $N_l = \{i: v_{il} = 1\}$, X_k, X_l, Y_k und Y_l jeweils Variablen, dann sind für die paarweisen Korrelationen zwischen zwei kardinal skalierten Merkmalen k und l, die sich allgemein nach der Formel

$$r_{kl}^{kardinal}(X_k, X_l, Y_k, Y_l) = \frac{\frac{1}{|N_{kl}|}\sum_{i \in N_{kl}}\left(a_{ik} - \frac{1}{|X_k|}\sum_{j \in X_k} a_{jk}\right)\left(a_{il} - \frac{1}{|X_l|}\sum_{j \in X_l} a_{jl}\right)}{\sqrt{\frac{1}{|Y_k|}\sum_{i \in Y_k}\left(a_{ik} - \frac{1}{|Y_k|}\sum_{j \in Y_k} a_{jk}\right)^2 \frac{1}{|Y_l|}\sum_{i \in Y_l}\left(a_{il} - \frac{1}{|Y_l|}\sum_{j \in Y_l} a_{jl}\right)^2}} \quad (4.3)$$

berechnen lassen, unter anderem die folgenden vier Varianten denkbar:

- $X_k = X_l = Y_k = Y_l = N_{kl}$, d.h. zur Berechnung der Kovarianzen und Varianzen werden die Mittelwerte über die paarweise vorhandenen Fälle bestimmt. Der Korrelationskoeffizient liegt immer im Intervall [-1;+1].
- $X_k = X_l = N_{kl}$, $Y_k = N_k$ und $Y_l = N_l$, d.h. zur Berechnung der Kovarianzen werden die Mittelwerte über die paarweise vorhandenen Fälle, zur Berechnung der Varianzen die Mittelwerte über die jeweils pro Merkmal vorhandenen Fälle bestimmt. Der Korrelationskoeffizient kann außerhalb des Intervalls [-1;+1] liegen.
- $X_k = N_k$, $X_l = N_l$ und $Y_k = Y_l = N_{kl}$, d.h. zur Berechnung der Kovarianzen werden die Mittelwerte über die jeweils pro Merkmal vorhandenen Fälle, zur Berechnung der Varianzen die Mittelwerte über die paarweise vorhandenen Fälle bestimmt. Der Korrelationskoeffizient kann außerhalb des Intervalls [-1;+1] liegen.
- $X_k = Y_k = N_k$ und $X_l = Y_l = N_l$, d.h. zur Berechnung der Kovarianzen und Varianzen werden die Mittelwerte über die jeweils pro Merkmal vorhandenen Fälle bestimmt. Der Korrelationskoeffizient kann außerhalb des Intervalls [-1;+1] liegen.

Die Tatsache, daß der Korrelationskoeffizient unter Umständen außerhalb des Intervalls [-1;+1] liegen kann, ist darauf zurückzuführen, daß in diesen Fällen unterschiedliche Stichprobenumfänge bei der Berechnung der Kovarianzen bzw. Varianzen herangezogen werden.

Beispiel:

Unter Verwendung der sechs kardinalen Merkmale Deskriptive Statistik (4), Testverfahren (5), Multivariate Verfahren (6), Spezialgebiete (7), Businessgrafiken (8) und Statistische Grafiken (9) der Datenmatrix des Anhangs A (Fall 1) ergeben sich beispielsweise die Mengen $N_4 = \{1,3,4,5,6,7,8,9,10,11,13,15\} = N_{4,5}$ und $N_5 = \{1,...,15\}$. Werden diese Mengen analog für die anderen Merkmale und Merkmalspaare bestimmt, dann können nach der Formel (4.3) unter Berücksichtigung der vier Varianten die folgenden Korrelationsmatrizen berechnet werden:

- $X_k = X_l = Y_k = Y_l = N_{kl} \; \forall \; k,l \in M \quad \Rightarrow \quad R = \begin{pmatrix} 1.000 & 0.156 & 0.802 & 0.666 & 0.355 & -0.020 \\ 0.156 & 1.000 & -0.134 & 0.064 & 0.090 & -0.262 \\ 0.802 & -0.134 & 1.000 & 0.820 & 0.607 & 0.498 \\ 0.666 & 0.064 & 0.820 & 1.000 & 0.766 & 0.687 \\ 0.355 & 0.090 & 0.607 & 0.766 & 1.000 & 0.710 \\ -0.020 & -0.262 & 0.498 & 0.687 & 0.710 & 1.000 \end{pmatrix}$.

- $X_k = X_l = N_{kl}$, $Y_k = N_k$ und $Y_l = N_l \; \forall \; k,l \in M \quad \Rightarrow \quad R = \begin{pmatrix} 1.000 & 0.167 & 0.845 & 0.629 & 0.374 & -0.022 \\ 0.167 & 1.000 & -0.134 & 0.070 & 0.090 & -0.278 \\ 0.845 & -0.134 & 1.000 & 0.854 & 0.607 & 0.514 \\ 0.629 & 0.070 & 0.854 & 1.000 & 0.798 & 0.652 \\ 0.374 & 0.090 & 0.607 & 0.798 & 1.000 & 0.697 \\ -0.022 & -0.278 & 0.514 & 0.652 & 0.697 & 1.000 \end{pmatrix}$.

- $X_k = N_k$, $X_l = N_l$ und $Y_k = Y_l = N_{kl} \; \forall \; k,l \in M \quad \Rightarrow \quad R = \begin{pmatrix} 1.000 & 0.143 & 0.735 & 0.596 & 0.325 & 0.015 \\ 0.143 & 1.000 & -0.134 & 0.058 & 0.090 & -0.242 \\ 0.735 & -0.134 & 1.000 & 0.752 & 0.607 & 0.459 \\ 0.596 & 0.058 & 0.752 & 1.000 & 0.702 & 0.602 \\ 0.325 & 0.090 & 0.607 & 0.702 & 1.000 & 0.656 \\ 0.015 & -0.242 & 0.459 & 0.602 & 0.656 & 1.000 \end{pmatrix}$.

- $X_k = Y_k = N_k$ und $X_l = Y_l = N_l \; \forall \; k,l \in M \quad \Rightarrow \quad R = \begin{pmatrix} 1.000 & 0.153 & 0.775 & 0.562 & 0.342 & 0.015 \\ 0.153 & 1.000 & -0.134 & 0.063 & 0.090 & -0.256 \\ 0.775 & -0.134 & 1.000 & 0.783 & 0.607 & 0.474 \\ 0.562 & 0.063 & 0.783 & 1.000 & 0.730 & 0.572 \\ 0.342 & 0.090 & 0.607 & 0.730 & 1.000 & 0.644 \\ 0.015 & -0.256 & 0.474 & 0.572 & 0.644 & 1.000 \end{pmatrix}$.

Werden die in den kardinalen Merkmalen der Datenmatrix als fehlend betrachteten Ausprägungen bei der Berechnung der Korrelationskoeffizienten verwendet, ergibt sich zum Vergleich die folgende „wahre" Korrelationsmatrix, wobei die Formel (4.3) mit $N_{kl} = N_k = N_l = N \; \forall \; k,l \in M$, d.h. die Datenmatrix ist vollständig, ebenfalls angewandt werden kann:

$$R = \begin{pmatrix} 1.000 & 0.122 & 0.669 & 0.479 & 0.336 & 0.079 \\ 0.122 & 1.000 & -0.134 & 0.081 & 0.090 & -0.290 \\ 0.669 & -0.134 & 1.000 & 0.761 & 0.607 & 0.541 \\ 0.479 & 0.081 & 0.761 & 1.000 & 0.739 & 0.624 \\ 0.336 & 0.090 & 0.607 & 0.739 & 1.000 & 0.718 \\ 0.079 & -0.290 & 0.541 & 0.624 & 0.718 & 1.000 \end{pmatrix}.$$

Analog können die Korrelationskoeffizienten auch bei Vorliegen eines anderen Skalenniveaus der Merkmale berechnet werden. Die paarweisen Korrelationen zwischen zwei ordinal skalierten Merkmalen k und l lassen sich beispielsweise nach der dem Rangkorrelationskoeffizienten von Spearman entsprechenden Formel

$$r_{kl}^{ordinal} = 1 - \frac{6 \cdot \sum_{i \in N_{kl}} (\rho_{ik} - \rho_{il})^2}{\left(|N_{kl}| - 1\right) \cdot |N_{kl}| \cdot \left(|N_{kl}| + 1\right)} \tag{4.4}$$

bestimmen. Dabei bezeichnet ρ_{ik} die Rangnummer, die der Ausprägung a_{ik} zugeordnet ist. Der nach (4.4) berechnete Korrelationskoeffizient liegt immer im Intervall [-1;+1] und ist identisch zu dem nach der Formel (4.3) mit $X_k = X_l = Y_k = Y_l = N_{kl}$ ermittelten Wert, wenn die Rangnummern den Merkmalsausprägungen entsprechen.

Die nach den Formeln (4.3) bzw. (4.4) berechnete Korrelationsmatrix R sowie eine entsprechend ermittelte Kovarianzmatrix S sind nicht notwendig positiv semidefinit (vgl. z.B. *Frane, 1978, S. 30, Heiberger, 1977, S. 40*). Dies kann bei einigen multivariaten Verfahren, wie beispielsweise der Faktoren- oder Diskriminanzanalyse, zu Problemen führen, da bei der Lösung der jeweiligen Optimierungsprobleme (vgl. z.B. *Hartung, Elpelt, 1992, S. 528* oder *Opitz, 1980, S. 116, 148*) negative Eigenwerte auftreten können (vgl. *Opitz, 1995, S. 372, Satz 6.35*). Darüber hinaus ist eine Invertierung von R, die beispielsweise zur Berechnung der Kleinst-Quadrate-Schätzwerte im Rahmen einer Regressionsanalyse nötig ist, nur dann möglich, wenn positive oder negative Definitheit vorliegt.[1] Dieses Problem ist jedoch unabhängig vom Vorliegen fehlender Werte in einer Datenmatrix immer gegeben, da selbst die Korrelationsmatrix einer vollständigen Datenmatrix notwendig nur positiv semidefinit ist.[2]

Für den Fall nicht positiv semidefiniter Korrelations- und Kovarianzmatrizen wird in der Literatur eine **Glättung** der entsprechenden Matrix vorgeschlagen, um so die Eigenschaft der positiven Semidefinitheit zu gewährleisten. Ein beispielsweise bei *Frane (1978, S. 30-31)* oder *Schwertman und Allen (1979, S. 187-188)* beschriebenes Verfahren der Glättung basiert auf der Approximationseigenschaft der **Singulärwertzerlegung** (singular value decomposition oder SVD), der **low rank matrix approximation**. Eine ausführliche Darstellung der Singulärwertzerlegung findet sich beispielsweise bei *Hauke (1992, S. 169-176)*. Für den Fall einer nicht positiv semidefiniten Korrelationsmatrix ergibt sich eine Glättung nach diesem Verfahren mittels der Formel

$$R^{geglättet} = F^{-1} \cdot U \cdot D_\lambda \cdot U^T \cdot F^{-1}, \tag{4.5}$$

wobei $D_\lambda \in \mathbb{R}^{p \times p}$ die Diagonalmatrix der positiven Eigenwerte $\lambda_1,...,\lambda_p$ der Korrelationsmatrix $R \in \mathbb{R}^{m \times m}$, $U \in \mathbb{R}^{m \times p}$ die Matrix der dazugehörigen Eigenvektoren und $F \in \mathbb{R}^{m \times m}$ die Diagonalmatrix der Quadratwurzeln der Diagonalelemente von $UD_\lambda U^T$, die lediglich eine normierende Wirkung besitzt, sind. Modifikationen dieses Ansatzes, wie beispielsweise die Anwendung auf eine Kovarianzmatrix oder der Verwendung von Eigenwerten ab einer vorgegebenen Größe, können analog durchgeführt werden.

---

[1] Für eine beliebige symmetrische Matrix $B \in \mathbb{R}^{n \times n}$ gilt:

B positiv oder negativ definit $\Rightarrow$ Eigenwerte $\lambda_1,...,\lambda_n \neq 0$

$\Leftrightarrow$ Rg B = n

$\Leftrightarrow$ B ist regulär.

[2] Sei $R = \tilde{A}^T\tilde{A}$ mit $\tilde{A} = (\tilde{a}_{ik})_{n,m}$, $\tilde{a}_{ik} = \frac{a_{ik} - \bar{a}_k}{\sqrt{s_{kk}}} \cdot \frac{1}{\sqrt{n}}$ $(i = 1,...,n,\ k = 1,...,m)$, wobei $\bar{a}_k = \frac{1}{n}\sum_{i=1}^{n} a_{ik}$ und $s_{kk} = \frac{1}{n}\sum_{i=1}^{n}(a_{ik} - \bar{a}_k)^2$, dann gilt: $x^T\tilde{A}^T\tilde{A}x = (\tilde{A}x)^T\tilde{A}x \geq 0$ für alle $x \neq 0$ $\Rightarrow$ $\tilde{A}^T\tilde{A} = R$ ist positiv semidefinit.

Als weitere Möglichkeit der Glättung wird von *Schnell (1986, S. 86)* die Anwendung einer sogenannten **Ridge-Prozedur** (RP) vorgeschlagen, bei der eine positive Zahl k bestimmt und zu jedem Diagonalelement der Korrelations- oder Kovarianzmatrix addiert wird. Zur Ermittlung der Zahl k erwähnt *Schnell* lediglich die Möglichkeit über Iterationen. Eine geglättete Korrelationsmatrix ergibt sich dann nach der Formel

$$R^{geglättet} = F^{-1} \cdot (R + k \cdot I) \cdot F^{-1}, \tag{4.6}$$

wobei die Diagonalmatrix $F \in \mathbb{R}^{m \times m}$, die aus den Quadratwurzeln der Diagonalelemente von $(R + kI)$ besteht, lediglich eine normierende Wirkung besitzt und $I \in \mathbb{R}^{m \times m}$ die Einheitsmatrix darstellt. Da sich die Eigenwerte der daraus resultierenden Matrix gerade um die addierte Zahl erhöhen,[3] sollte diese nicht iterativ sondern deterministisch gleich dem Betrag des ursprünglich kleinsten, negativen Eigenwerts gewählt werden.

Die Glättung einer nicht positiv semidefiniten Korrelations- oder Kovarianzmatrix mittels der Ridge-Prozedur besitzt im Vergleich zur Glättung auf Basis der Singulärwertzerlegung den Vorteil der einfacheren Berechnung. Dafür führt die Ridge-Prozedur zu einer Verzerrung der berechneten Koeffizienten (*Freud, Minton, 1979, S. 140*), während die Anwendung der Singulärwertzerlegung eine bestmögliche Approximation der Ausgangsmatrix im Sinne eines Kleinst-Quadrate-Kriteriums darstellt (vgl. *Hauke, 1992, S. 173*).

Beispiel:

Für die aus den kardinalen Merkmalen der Datenmatrix des Anhangs A (Fall 1) berechnete, nicht positiv semidefinite Korrelationsmatrix des Falls $X_k = X_l = N_{kl}$, $Y_k = N_k$, $Y_l = N_l \; \forall \; k,l \in M$ (die Eigenwerte betragen 3.46, 1.35, 0.92, 0.21, 0.09 und -0.01) ergeben sich unter Anwendung der beiden Glättungsmethoden die folgenden Ergebnisse, wobei λ den Vektor der Eigenwerte darstellt:

- Glättung auf Basis der SVD:

$$R^{geglättet} = \begin{pmatrix} 1.000 & 0.166 & 0.836 & 0.628 & 0.372 & -0.019 \\ 0.166 & 1.000 & -0.133 & 0.070 & 0.090 & -0.278 \\ 0.836 & -0.133 & 1.000 & 0.851 & 0.606 & 0.509 \\ 0.628 & 0.070 & 0.851 & 1.000 & 0.798 & 0.652 \\ 0.372 & 0.090 & 0.606 & 0.798 & 1.000 & 0.696 \\ -0.019 & -0.278 & 0.509 & 0.652 & 0.696 & 1.000 \end{pmatrix}, \lambda = \begin{pmatrix} 3.42 \\ 1.34 \\ 0.92 \\ 0.22 \\ 0.09 \\ 0.00 \end{pmatrix}.$$

- Glättung mittels der RP:

$$R^{geglättet} = \begin{pmatrix} 1.000 & 0.165 & 0.835 & 0.622 & 0.370 & -0.022 \\ 0.165 & 1.000 & -0.132 & 0.069 & 0.089 & -0.275 \\ 0.835 & -0.132 & 1.000 & 0.844 & 0.600 & 0.508 \\ 0.622 & 0.069 & 0.844 & 1.000 & 0.789 & 0.645 \\ 0.370 & 0.089 & 0.600 & 0.789 & 1.000 & 0.689 \\ -0.022 & -0.275 & 0.508 & 0.645 & 0.689 & 1.000 \end{pmatrix}, \lambda = \begin{pmatrix} 3.44 \\ 1.35 \\ 0.92 \\ 0.21 \\ 0.09 \\ 0.00 \end{pmatrix}.$$

Bei den anderen drei Varianten der Formel (4.3) sind die im Beispiel berechneten Korrelationsmatrizen bereits jeweils positiv definit, so daß ein Glätten dieser Matrizen entfällt.

---

[3] Sei $B \in \mathbb{R}^{m \times m}$ eine symmetrische Matrix, $D_\lambda \in \mathbb{R}^{m \times m}$ die Diagonalmatrix der Eigenwerte, $U \in \mathbb{R}^{m \times m}$ die Matrix der Eigenvektoren, $I \in \mathbb{R}^{m \times m}$ die Einheitsmatrix sowie k eine positive reelle Zahl. Dann gilt:

$B + kI = UD_\lambda U^T + kI \Leftrightarrow (B + kI)U = (UD_\lambda U^T + kI)U = UD_\lambda U^T U + (kI)U = UD_\lambda + U(kI) = U(D_\lambda + kI)$

4.1.2 Merkmalseliminierung

Eine weitere Eliminierungsstrategie, die im Vergleich zu den bisher vorgestellten Ansätzen der Objekteliminierung zwar inhaltlich verschieden, aber formal gleich ist, stellt der Ausschluß von Merkmalen mit fehlenden Werten aus der Untersuchung, die sogenannte Merkmalseliminierung, dar. In der Literatur erfolgt eine Behandlung des Problems fehlender Daten in erster Linie vor dem Hintergrund einer Analyse der Merkmale, d.h. auf Basis der unvollständigen Datenmatrix sollen Aussagen über die erhobenen Merkmale getroffen werden. Deshalb existieren bislang kaum Arbeiten, in denen das Entfernen von Merkmalen dargestellt und untersucht wird.

4.1.2.1 Analyse der vollständigen Merkmale

Werden bei einer Auswertung nur die Merkmale verwendet, deren Ausprägungen bezüglich aller untersuchten Objekte vorliegen, dann spricht man von einer Analyse der vollständig erhobenen Merkmale (**complete-variable analysis**). Wird die Datenmatrix A in der Form

$$A = (a_{ik})_{n,m} = (A_{obs}, A_{mis}) = \left((a_{ik})_{n,q} , (a_{ik})_{n,m-q} \right) \tag{4.7}$$

geschrieben, wobei A_{obs} die q vollständigen und A_{mis} die $(m - q)$ unvollständigen Merkmalsvektoren enthält, dann wird bei einer complete-variable analysis lediglich die Matrix A_{obs} verwendet, die im Vergleich zur Matrix A die niedrigere Dimension $(n \times q)$ besitzt.

Beispiel:

Für die Datenmatrix des Anhangs A (Fall 1) ergibt sich nach diesem Verfahren eine auswertbare Matrix A_{obs} der Dimension (15 × 3), wobei die Merkmale Preisniveau, Benutzeroberfläche, Programmierbarkeit, Deskriptive Statistik, Spezialgebiete und Statistische Grafiken aus der weiteren Analyse eliminiert werden.

Da in der Datenanalyse eine Untersuchung der Ähnlichkeitsstruktur der Objekte von Interesse ist, stellt das Entfernen einzelner Merkmale im Vergleich zur Objekteliminierung eine geeignetere Lösung im Hinblick auf das Analyseziel dar (vgl. *Wishart, 1978, S. 282*). Dabei ist jedoch zu beachten, daß die Merkmale im allgemeinen nicht grundlos für die Untersuchung ausgewählt und erhoben wurden (vgl. *Evans et al., 1979, S. 469*). Bei einer complete-variable analysis werden zwar alle erhobenen Objekte in der weiteren Analyse verwendet, jedoch kann der Fall eintreten, daß die Objekte durch die verbleibenden Merkmale nicht mehr angemessen beschrieben werden.

4.1.2.2 Analyse der verfügbaren Merkmale

Analog zu Abschnitt 4.1.1.2 ist auch das Verwenden der jeweils verfügbaren Merkmale in einer Analyse (**available-variable analysis**) denkbar. In der Datenanalyse sind die folgenden zwei Möglichkeiten einer derartigen Auswertung von Interesse:

- Aus einer zuvor standardisierten Datenmatrix werden die paarweisen Korrelationskoeffizienten zwischen jeweils zwei Objekten berechnet. Die auf diese Art erhaltene Korrelationsmatrix kann dann beispielsweise in eine Distanzmatrix transformiert werden, indem zunächst zu jeder Komponente der Korrelationsmatrix die Zahl (–1) addiert und anschließend die sich dadurch ergebende Matrix mit $(-a)$ multipliziert wird, wobei a eine positive reelle Zahl ist. Die Werte der resultierenden Distanzmatrix liegen dann im Intervall $[0;2a]$.
- Mit den paarweise verfügbaren Merkmalen werden die Distanzindizes zwischen jeweils zwei Objekten unmittelbar bestimmt und in einer Distanzmatrix zusammengefaßt.

In beiden Fällen handelt es sich genau genommen um die sogenannte **pairwise available-variable analysis**, d.h. es werden nur die Merkmale herangezogen, deren Ausprägungen für die beiden jeweils betrachteten Objekte vorhandenen sind. Im ersten Fall kann eine Berechnung der paarweisen Korrelationen zwischen zwei Objekten in Anlehnung an die Formel (4.3) erfolgen. Da die aus einer Korrelationsmatrix der Objekte ermittelte Distanzmatrix eher die Interdependenzen als die Ähnlichkeiten der Objektpaare widerspiegelt, soll in diesem Abschnitt lediglich die unmittelbare Berechnung einer Distanzmatrix ausführlicher dargestellt werden.

Ausgangspunkt dieses Ansatzes sind die jeweils für zwei Objekte $i, j \in N$ verfügbaren Merkmale, die in einer Menge $M_{ij} = \{k: v_{ik} = 1 \wedge v_{jk} = 1\}$ zusammengefaßt werden. Zur Bestimmung einer Distanz zwischen den beiden Objekten i und j werden dann die Merkmale $k \in M_{ij}$ herangezogen, also alle Merkmale, deren Ausprägungen bei beiden Objekten vorliegen. Zusätzlich erfolgt eine Normierung unter Verwendung der Anzahl der verfügbaren Merkmale $|M_{ij}|$ im Verhältnis zur Elementanzahl $|M|$ der Merkmalsmenge. Für eine unvollständige quantitative Datenmatrix kann damit die Verschiedenheit zweier Objekte $i, j \in N$ beispielsweise durch eine auf diese Art modifizierte **gewichtete L_p-Distanz**

$$d_{ij} = \left(\frac{|M|}{|M_{ij}|} \sum_{k \in M_{ij}} \alpha_k \cdot |a_{ik} - a_{jk}|^p \right)^{\frac{1}{p}} \qquad (\alpha_k \geq 0 \; \forall \; k \in M) \tag{4.8}$$

gemessen werden. Die nichtnegativen reellen Zahlen $\alpha_1,...,\alpha_m$ sind merkmalsspezifische lineare Gewichtungen und der Parameter p bestimmt das Ausmaß, wie stark die absolu-

ten Differenzen der Merkmalsausprägungen berücksichtigt werden. Im Fall einer unvollständigen gemischten Datenmatrix kann z.B. in Anlehnung an *Wishart (1985, S. 126-128 und 1986, S. 454-457)* die **linearhomogene Aggregation** merkmalsweiser Distanzindizes folgendermaßen modifiziert werden:

$$d_{ij} = \frac{|M|}{|M_{ij}|} \sum_{k \in M_{ij}} \alpha_k \cdot d_k(i,j) \qquad (\alpha_k \geq 0 \; \forall \; k \in M). \tag{4.9}$$

Die merkmalsweisen Distanzen $d_k(i,j)$ können dabei entsprechend dem Skalenniveau der Merkmale wie im Falle einer vollständigen Datenmatrix bestimmt werden (vgl. z.B. *Ambrosi, 1981, S. 8-39* oder *Opitz, 1980, S. 32-50*). Die nichtnegativen reellen Zahlen $\alpha_1,...,\alpha_m$ sind wiederum merkmalsspezifische Gewichte und dienen in erster Linie dazu, die möglicherweise vorhandenen, unterschiedlichen Schwankungsbreiten der merkmalsweisen Distanzen auszugleichen. Analog sind Modifikationen weiterer, in der Literatur bekannter Distanzindizes (vgl. z.B. *Opitz, 1980, S. 50-64* oder *Sodeur, 1974, S. 80-111*) denkbar, wobei an dieser Stelle auf eine detaillierte Darstellung verzichtet wird.

Beispiel:

Für die Objekte CRUNCH (2), MICROSTAT (4) und STATISTIX (13) der Datenmatrix des Anhangs A (Fall 1) ergeben sich gemäß der Formel (4.9) beispielsweise die paarweisen Distanzen $d_{24} = 1.44$, $d_{2,13} = 2.49$ und $d_{4,13} = 0.62$, wobei der Berechnung folgendes zugrundegelegt wurde:

- $|M_{24}| = 4, |M_{2,13}| = 7, |M_{4,13}| = 6, \; \alpha_k = \dfrac{1}{\max\limits_{i,j \in N_k} d_k(i,j)} \; \forall \; k \in M \quad (N_k = \{i: v_{ik} = 1\})$.
- Merkmal k nominal: $d_k(i,j) = \begin{cases} 0 & a_{ik} = a_{jk} \\ 1 & a_{ik} \neq a_{jk} \end{cases} \quad (i,j \in N)$.
- Merkmal k ordinal: $d_k(i,j) = \left| (\text{Rang}\, a_{ik}) - (\text{Rang}\, a_{jk}) \right| \quad (i,j \in N)$.
- Merkmal k kardinal: $d_k(i,j) = |a_{ik} - a_{jk}| \quad (i,j \in N)$.

Werden die in der Datenmatrix als fehlend betrachteten Ausprägungen bei der Berechnung verwendet, ergeben sich zum Vergleich die Distanzindizes $d_{24} = 1.62$, $d_{2,13} = 2.74$ und $d_{4,13} = 1.51$. In diesem Fall ist in der Formel (4.9) $M_{ij} = M \; \forall \; i,j \in N$ zu setzen.

Analog zum Problem in Abschnitt 4.1.1.2, daß eine nach den Formeln (4.3) bzw. (4.4) berechnete Korrelationsmatrix oder eine entsprechend ermittelte Kovarianzmatrix nicht notwendig positiv semidefinit ist, müssen die nach den Formeln (4.8) oder (4.9) ermittelten Distanzindizes im Fall fehlender Daten nicht notwendig die Eigenschaft einer Quasimetrik besitzen (vgl. *Gower, 1971, S. 860-861, 870-871*). Die Eigenschaft einer Quasimetrik für einen Distanzindex ist zwar im Vergleich zur Eigenschaft der positiven Semidefinitheit einer Korrelationsmatrix nicht von entscheidender Bedeutung, jedoch sollen mit der nachfolgenden Darstellung Parallelen zum Abschnitt 4.1.1.2 aufgezeigt werden.

Der nach (4.8) oder (4.9) im Fall fehlender Daten ermittelte Distanzindex ist keine Quasimetrik, sobald die Dreiecksungleichung

$$h,i,j \in N \quad \Rightarrow \quad d_{ij} \leq d_{ih} + d_{jh} \tag{4.10}$$

verletzt ist. Während im Fall einer vollständigen Datenmatrix der nach (4.8) ermittelte Distanzindex der Ungleichung (4.10) genügt, also eine Quasimetrik ist, besitzt ein gemäß (4.9) bestimmter Distanzindex diese Eigenschaft nur dann, wenn die merkmalsweisen Distanzindizes ebenfalls mindestens eine Quasimetrik sind.

Beispiel:

Für die oben berechneten Distanzindizes zwischen den Objekten CRUNCH, MICROSTAT und STATISTIX im Fall fehlender Daten ist die Dreiecksungleichung (4.10) nicht erfüllt, da beispielsweise gilt:

$$d_{2,13} = 2.49 \quad > \quad d_{24} + d_{4,13} = 1.44 + 0.62 = 2.06 .$$

Dagegen erfüllen die im Fall ohne fehlende Daten berechneten Distanzindizes d_{24} = 1.62, $d_{2,13}$ = 2.74 und $d_{4,13}$ = 1.51 die Ungleichung (4.10), da die verwendeten merkmalsweisen Distanzindizes mindestens die Eigenschaft einer Quasimetrik besitzen.

Grundsätzlich ist eine Glättung der aus einer unvollständigen Datenmatrix berechneten Distanzmatrix denkbar, so daß schließlich die Ungleichung (4.10) für alle paarweisen Distanzindizes erfüllt ist. In Anlehnung an *Bock (1974, S. 373-374)* wird dazu in allen nicht erfüllten Dreiecksungleichungen der Distanzindex d_{ij} gleich der Summe von d_{ih} und d_{jh} gesetzt. Dadurch können vorher gültige Dreiecksungleichungen verletzt werden. Das Verfahren muß deshalb iterativ solange angewandt werden, bis (4.10) für alle paarweisen Distanzindizes erfüllt ist. Aufgrund der endlich vielen Werte in der Distanzmatrix bricht das Verfahren nach endlich vielen Schritten ab.

Beispiel:

Der Distanzindex $d_{2,13}$ wird gleich der Summe von d_{24} und $d_{4,13}$ gesetzt, d.h. $d_{2,13}$ = 1.44 + 0.62 = 2.06, so daß die Ungleichung $d_{2,13}$ = 2.06 ≤ d_{24} + $d_{4,13}$ = 1.44 + 0.62 = 2.06 erfüllt ist. Da alle möglichen Dreiecksungleichungen mit $d_{2,13}$ auf der rechten Seite des Ungleichheitszeichens trotz des geänderten Werts für diesen Distanzindex gültig bleiben, sind Anpassungen weiterer Distanzindizes nicht nötig.

Ein weiterer Ansatz zur Glättung der aus einer unvollständigen Datenmatrix berechneten Distanzmatrix besteht darin, eine Konstante zu allen paarweisen Distanzindizes zu addieren, so daß die Dreiecksungleichung für alle paarweisen Distanzindizes erfüllt ist. In der Literatur wird in diesem Zusammenhang von einem sogenannten **additive constant problem** gesprochen (vgl. z.B. *Furnas, 1989, S. 17, Cailliez, 1983*). Grundsätzlich existieren unendlich viele Lösungen für die Wahl einer derartigen Konstanten. Das Minimum dieser Lösungsmenge und damit der kleinste Wert, den die zu addierende Kon-

stante annehmen muß, damit die Dreiecksungleichung für alle paarweisen Distanzindizes gerade erfüllt ist, ergibt sich gemäß

$$\max_{i,j,h \in N} \{d_{ij} - d_{ih} - d_{jh}\}. \tag{4.11}$$

Erfüllen alle paarweisen Distanzindizes die Dreiecksungleichung, dann ist der nach (4.11) ermittelte Wert kleiner gleich Null. Im Fall eines negativen Werts für die Konstante wäre also trotz der damit verbundenen Verkleinerung aller paarweisen Distanzindizes die Dreiecksungleichung immer noch erfüllt. Von Interesse ist jedoch ein sich nach (4.11) ergebender positiver Wert, da in diesem Fall die Ungleichung (4.10) zunächst nicht erfüllt ist. Durch die Addition dieses positiven Werts zu allen paarweisen Distanzen wird die Gültigkeit der Dreiecksungleichung jedoch gewährleistet.

Beispiel:

Für die Distanzindizes zwischen den Objekten CRUNCH, MICROSTAT und STATISTIX erhält man als zu addierende Konstante nach (4.11) einen Wert von 0.43. Die entsprechend transformierten Distanzindizes ergeben sich dann mit $d_{24} = 1.87$, $d_{2,13} = 2.92$ und $d_{4,13} = 1.05$.

4.1.3 Vergleich und Kombination der Verfahren

In der nachfolgenden Tabelle 4.1 sind die Voraussetzungen sowie die wichtigsten Eigenschaften der vier vorgestellten Eliminierungsansätze noch einmal zusammenfassend dargestellt:

Verfahren	Voraussetzungen	Eigenschaften
Complete-case analysis	Daten- oder Distanzmatrix, Daten sind MCAR	Unter Umständen erheblicher Informationsverlust, Analyse der Ähnlichkeitsstruktur aller Objekte nicht mehr möglich
Available-case analysis	Datenmatrix, Daten sind MCAR	Kein bzw. geringer Informationsverlust, variierender Stichprobenumfang, Korrelations- bzw. Kovarianzmatrix nicht notwendig positiv semidefinit
Complete-variable analysis	Datenmatrix, Daten sind MCAR	Unter Umständen erheblicher Informationsverlust, Analyse der Ähnlichkeitsstruktur aller Objekte mit den verbleibenden Merkmalen möglich
Available-variable analysis	Datenmatrix, Daten sind MCAR	Kein bzw. geringer Informationsverlust, Ergebnis ist eine vollständige, mit geeigneten Methoden auswertbare Distanzmatrix

Tabelle 4.1: Vergleich der Eliminierungsverfahren

Die Eignung der beschriebenen Eliminierungsverfahren im Rahmen einer Datenanalyse ist zum einen abhängig von der Anzahl und den Positionen der fehlenden Werte in der Daten- bzw. Distanzmatrix. So ist beispielsweise bei einer Datenmatrix, in der lediglich ein Merkmal einen sehr hohen Anteil fehlender Werte aufweist, das Entfernen dieses Merkmals zweckmäßiger als das Eliminieren aller unvollständig erhobenen Objekte. Zum anderen ist das jeweils Untersuchungsziel für die Eignung einer Eliminierungstechnik von Bedeutung. Bei einer Analyse der Ähnlichkeitsstruktur der Objekte wird das Entfernen von Merkmalen aus der weiteren Analyse unter Umständen eine geeignetere Lösung darstellen als das Eliminieren der Objekte mit fehlenden Daten. Um jedoch Aussagen über die erhobenen Merkmale treffen zu können, ist eine Eliminierung aller Merkmale mit fehlenden Werten sicherlich keine zweckmäßige Vorgehensweise.

Darüber hinaus kann eine Kombination unterschiedlicher Eliminierungstechniken, wie sie beispielsweise von *Dempster (1971, S. 343)* vorgeschlagen wird, eine befriedigendere Lösung darstellen als die Anwendung eines einzigen Verfahrens. Einen derartigen Ansatz stellt das folgende Optimierungsproblem dar:

$$\text{Informationsverlust} \rightarrow \min \quad \Leftrightarrow \quad (o_1+\ldots+o_n)\cdot(m_1+\ldots+m_m)=\sum_{i=1}^{n}\sum_{k=1}^{m} o_i \cdot m_k \rightarrow \max \tag{4.12}$$

$$\text{mit} \quad \begin{pmatrix} o_1 \\ \vdots \\ o_n \end{pmatrix}(m_1,\ldots,m_m) \leq \begin{pmatrix} v_{11} & \ldots & v_{1m} \\ \vdots & & \vdots \\ v_{n1} & \ldots & v_{nm} \end{pmatrix} \quad \Leftrightarrow \quad o \cdot m^T \leq V,$$

$$o_i = \begin{cases} 1 & \text{falls Objekt } i \text{ verwendet wird} \\ 0 & \text{sonst,} \end{cases}$$

$$m_k = \begin{cases} 1 & \text{falls Merkmal } k \text{ verwendet wird} \\ 0 & \text{sonst,} \end{cases}$$

$$v_{ik} = \begin{cases} 1 & \text{falls Ausprägung } a_{ik} \text{ vorhanden} \\ 0 & \text{sonst.} \end{cases}$$

Mit p aus (4.1) und q aus (4.7) sowie o^* und m^* als optimale Lösungen von (4.12) gilt:

$$o_i^* = 1 \,\forall\, i\colon M_i = \{k\colon v_{ik} = 1\} = M,\ m_k^* = 1 \,\forall\, k\colon N_k = \{i\colon v_{ik} = 1\} = N, \tag{4.13}$$

$$\sum_{i=1}^{n}\sum_{k=1}^{m} o_i^* \cdot m_k^* \geq \max\{(p \cdot m),(n \cdot q)\}.$$

Die Lösung des Optimierungsproblems (4.12) kann beispielsweise mit Hilfe des allgemeinen **Branch and Bound Prinzips** (vgl. z.B. *Neumann, 1975, S. 333-339*) erfolgen. Mit (4.13) ergibt sich neben einer Eingrenzung der Menge der Optimallösungen bereits eine erste untere Schranke für den Zielfunktionswert.

Beispiel:

Für die Datenmatrix des Anhangs A (Fall 1) ergibt sich nach (4.13):

$o_i^* = 1 \,\forall\, i \in \{1,3,5,6,7,8,13\}, m_k^* = 1 \,\forall\, k \in \{5,6,8\}, \sum_{i=1}^{n}\sum_{k=1}^{m} o_i^* \cdot m_k^* \geq 63$.

Als optimale Lösung von (4.12) erhält man schließlich:

$o^{T^*} = (1,1,1,0,1,1,1,1,0,1,0,1,1,1,1)$, $m^{T^*} = (0,1,1,0,1,1,0,1,1)$ und $\sum_{i=1}^{n}\sum_{k=1}^{m} o_i^* \cdot m_k^* = 72$,

d.h. die Objekte MICROSTAT, SAS und STATA sowie die Merkmale Preisniveau, Deskriptive Statistik und Spezialgebiete werden eliminiert, so daß eine Datenmatrix der Dimension (12 × 6) mit 72 Ausprägungen für eine weitere Analyse zur Verfügung steht. Gegenüber den Lösungen von (4.1) mit 7·9 = 63 bzw. von (4.7) mit 15·3 = 45 verbleibenden Ausprägungen ergibt sich damit eine deutliche Verringerung des Informationsverlusts. Dennoch stehen für die weitere Analyse lediglich 12 Objekte und sechs Merkmale zur Verfügung, so daß eine Eliminierungsstrategie in diesem Fall nicht als geeignete Lösung des Problems fehlender Daten angesehen werden kann und die in den folgenden Kapiteln betrachteten Strategien unter Umständen diesem Lösungsansatz vorzuziehen sind.

4.2 Imputationsverfahren

Mit den Imputationsverfahren werden die bereits vorhandenen Daten um Schätzungen für die fehlenden Werte ergänzt, so daß eine vollständige Daten- oder Distanzmatrix gleicher Dimension resultiert. In der Literatur wird der Begriff **Ersetzungsverfahren** synonym zur Bezeichnung Imputationsverfahren verwendet, da im Prinzip die fehlenden Daten durch entsprechende Schätzwerte ersetzt werden. Darüber hinaus erscheinen jedoch auch die Begriffe **Ergänzungs-** oder **Vervollständigungsverfahren** zweckmäßig, da die vorhandenen Daten um Schätzwerte für die fehlenden Werte ergänzt oder vervollständigt werden.

Der Hauptvorteil der Imputationsverfahren liegt darin, daß die damit erzeugte vollständige Daten- oder Distanzmatrix, wie im Fall ohne fehlende Werte, mit einem entsprechend der Zielsetzung der Analyse geeigneten Verfahren ausgewertet werden kann und eine Eliminierung eines unter Umständen nicht mehr tolerierbaren Teils der Daten nicht in Kauf genommen werden muß (vgl. *Schnell, 1985, S. 92-93*). Darüber hinaus stellt die Verwendung der Informationen, die in den vorhandenen Daten stecken, eine zweckmäßige Vorgehensweise zur Schätzung der fehlenden Werte dar (vgl. *Chapman, 1976, S. 245*). Argumente gegen die Anwendung von Imputationstechniken beziehen sich vor allem auf die dadurch bedingten, möglichen Verzerrungen der Analyseergebnisse. Es wird sich jedoch zeigen, daß die Imputationsverfahren grundsätzlich nur unter gewissen Annahmen über den Ausfallmechanismus geeignet sind, d.h. abhängig von der Ersetzungstechnik müssen die Daten MCAR, MAR oder MCARC sein. Bei entsprechender Beachtung dieser Annahmen können die Verzerrungen der Analyseergebnisse ver-

mieden oder zumindest auf ein akzeptables Ausmaß reduziert werden. Dennoch ist zu beachten, daß auch im Fall unverzerrter Analyseergebnisse die Imputationswerte lediglich Schätzungen für die fehlenden Daten und nicht die tatsächlichen Werte darstellen (vgl. *Madow et al., 1983, S. 86-88*).

Die in der Literatur bekannten Imputationsverfahren gehen grundsätzlich von einer unvollständigen Datenmatrix A aus und liefern entsprechende Schätzungen für die fehlenden Ausprägungen a_{ik} $(i \in N, k \in M)$. Im Rahmen dieses Abschnittes sollen zusätzlich die denkbaren Modifikationen der einzelnen Verfahren für eine Imputation der fehlenden paarweisen Distanzen d_{ij} $(i, j \in N)$ einer unvollständigen, unmittelbar erhobenen Distanzmatrix D diskutiert werden.

In **Abschnitt 4.2.1** werden zunächst einfache Imputationstechniken, die auf der MCAR-Annahme beruhen, dargestellt. Dazu zählen die Imputation des Lageparameters oder des Verhältnisschätzers sowie die Imputation mittels Zufallsauswahl oder auf Basis von Expertenratings. Der **Abschnitt 4.2.2** widmet sich dann der Imputation innerhalb von Klassen. Diese Verfahren basieren auf dem gegenüber der MCAR-Annahme weniger restriktiven MCARC-Modell. Dabei werden zum einen die Möglichkeiten der Bestimmung geeigneter Klassen sowie die auf diesen Klassen basierenden Imputationsverfahren, zu denen in erster Linie die Cold-Deck- und Hot-Deck-Techniken zählen, dargestellt. Der **Abschnitt 4.2.3** behandelt anschließend die multivariaten Imputationstechniken, die zwar grundsätzlich die MCAR-Annahme benötigen, aber in bestimmten Fällen auch mit der MAR-Annahme auskommen. Bei Vorliegen einer unvollständigen Datenmatrix kann die Imputation mittels der Regressions-, Varianz-, Diskriminanz- oder Hauptkomponentenanalyse erfolgen. Mit diesen Analysemethoden können die Zusammenhänge zwischen den Merkmalen aufgedeckt und zur Bestimmung der Imputationswerte genutzt werden. Im Fall einer unvollständigen, unmittelbar erhobenen Distanzmatrix ist eine Imputation unter Verwendung von Distanzeigenschaften möglich. Als Distanzeigenschaften werden die Dreiecksungleichung und die ultrametrische Ungleichung betrachtet. Obwohl deren Gültigkeit für die Datenanalyse nicht von entscheidender Bedeutung ist, stellen diese Eigenschaften eine einfache Möglichkeit dar, Abhängigkeiten zwischen den paarweisen Distanzen zu berücksichtigen und für die Bestimmung von Imputationswerten zu nutzen. Damit ist auch die Einordnung dieses Ansatzes unter dem Begriff der multivariaten Imputationstechniken zu begründen. In **Abschnitt 4.2.4** werden dann noch kurz die Möglichkeiten einer Imputation bei systematischen Ausfallmechanismen beschrieben. Die wenigen, im Fall nicht zufällig fehlender Daten geeigneten Ansätze ziehen zur Bestimmung der Imputationswerte ein Modell des zugrundeliegenden Ausfallmechanismus heran. In **Abschnitt 4.2.5** erfolgt abschließend ein Vergleich der vorgestellten Imputationsverfahren sowie eine Diskussion der Kombinationsmöglichkeiten einzelner Techniken.

4.2.1 Einfache Imputationstechniken

Die hier dargestellten Imputationstechniken liefern ohne großen Aufwand Schätzungen für die fehlenden Daten und setzen voraus, daß das Fehlen einzelner Werte in keiner Beziehung zu den fehlenden sowie den vorhandenen Werten der Daten- oder Distanzmatrix steht. Dabei wird die Möglichkeit der Ersetzung eines unvollständigen Objektvektors durch einen vollständigen Objektvektor, wie dies beispielsweise von *Little und Rubin (1987, S. 60)* oder *Chapman (1976, S. 250, 1983b, S. 46-48)* vorgeschlagen wird, nicht weiter betrachtet, da dieser Ansatz prinzipiell einer complete-case analysis unter Hinzunahme neu erhobener Objekte entspricht.

4.2.1.1 Imputation des Lageparameter

Die Ausprägungen eines Merkmals können durch eine einzige Zahl, einen sogenannten Lageparameter, charakterisiert werden. Dieser Lageparameter kann als Imputationswert für die fehlenden Daten des Merkmals herangezogen werden. Abhängig vom Skalenniveau des Merkmals stellen somit der Modus, der Median und das arithmetische Mittel mögliche Imputationswerte dar. Die Bestimmung von Lageparametern kann sowohl rein deskriptiv wie auch unter der Annahme einer speziellen Verteilung der als Zufallsvariablen betrachteten Merkmale erfolgen, wobei im letzteren Fall der Erwartungswert anstelle des arithmetischen Mittels zu verwenden ist (vgl. *Bamberg, Baur, 1992, S. 16, S. 119-120*).

Der erste Ansatz einer Imputation des Lageparameters geht auf *Wilks (1932)* zurück. In seiner Arbeit verwendet *Wilks* das **arithmetische Mittel** als Schätzwert für die fehlenden Daten. An dieser Stelle ist jedoch anzumerken, daß *Wilks* sich eigentlich mit der Schätzung der Erwartungswerte sowie der Kovarianzmatrix im bivariaten Fall unter der Annahme einer Normalverteilung, also einem Parameterschätzproblem, beschäftigte. Mit $N_k = \{i: v_{ik} = 1\}$ können in Anlehnung an *Toutenburg (1992, S. 210)* die Schätzwerte für die fehlenden Ausprägungen kardinal skalierter Merkmale einer Datenmatrix gemäß der Formel

$$a_{ik} = \bar{a}_k = \frac{1}{|N_k|} \sum_{j \in N_k} a_{jk} \quad \forall\, i,k: v_{ik} = 0 \tag{4.14}$$

berechnet werden. Der Mittelwert entspricht unter der Annahme einer Normalverteilung einer Maximum-Likelihood-Schätzung und ohne Normalverteilungsannahme nach der Kleinst-Quadrate-Methode einem Schätzwert mit minimalem Fehler für die nicht vorhandenen Daten (vgl. *Lösel, Wüstendörfer, 1974, S. 348*). Damit ist auch der in der Literatur häufig verwendete Begriff der **Zero-Order-Regression** zu erklären. Durch

die Mittelwertersetzung ändert sich zwar das arithmetische Mittel des vervollständigten Merkmals nicht, jedoch muß im Regelfall eine Unterschätzung der tatsächlichen Varianz in Kauf genommen werden (vgl. *Little, Rubin, 1987, S. 44*).

Im Fall fehlender Ausprägungen bei nominalen bzw. ordinalen Merkmalen kann als Imputationswert der **Modus** bzw. der **Median** herangezogen werden, d.h.

$$a_{ik} = a_k^{Mod} \quad \forall\, i,k: v_{ik} = 0 \quad (k \text{ nominal}), \tag{4.15}$$

$$a_{ik} = a_k^{Med} \quad \forall\, i,k: v_{ik} = 0 \quad (k \text{ ordinal}), \tag{4.16}$$

wobei eine Definition der beiden Lageparameter beispielsweise bei *Bamberg und Baur (1992, S. 16-17)* zu finden ist.

Beispiel:

Für die fehlenden Ausprägungen der Datenmatrix des Anhangs A (Fall 1) ergeben sich nach (4.14), (4.15) und (4.16) die folgenden Imputationswerte:

Fehlende Ausprägung	$a_{21}, a_{10,1}$	a_{42}	a_{43}, a_{93}	$a_{24}, a_{12,4}, a_{14,4}$	$a_{12,7}, a_{14,7}, a_{15,7}$	$a_{49}, a_{11,9}$
Verwendeter Lageparameter	a_1^{Med}	a_2^{Mod}	a_3^{Mod}	$\bar{a}_4$	$\bar{a}_7$	$\bar{a}_9$
Imputationswert	mittel	M/K	ja	76.61	44.47	40.15

Bei einer unmittelbar erhobenen Distanzmatrix D kann ein zur Imputation des Lageparameters analoges Verfahren angewandt werden. Dabei wird eine fehlende Distanz der Objekte $i, j \in N$ durch den Mittelwert der vorhandenen Distanzen der Objekte i und j mit den jeweils anderen Objekten $h \in N, h \neq i$ bzw. $h \neq j$ geschätzt, d.h.

$$d_{ij} = \frac{1}{|N^i| + |N^j|} \left(\sum_{h \in N^i} d_{ih} + \sum_{h \in N^j} d_{jh} \right) \quad \forall\, i,j: w_{ij} = 0. \tag{4.17}$$

Dabei stellt $N^i = \{h: w_{ih} = 1, h \neq i\}$ die Menge von Objekten, deren Distanzen bezüglich des Objekts i vorhanden sind, dar. Anstelle des Mittelwerts ist auch der Median ein bei Distanzen geeigneter Lageparameter zur Schätzung der fehlenden Werte.

Beispiel:

Für die fehlenden paarweisen Distanzen der Distanzmatrix des Anhangs B (Fall 1) ergeben sich nach (4.17) die folgenden Imputationswerte:

Fehlende Distanz	d_{14}	$d_{1,10}$	d_{27}	$d_{2,10}$	d_{35}	d_{37}	d_{46}	d_{59}	d_{69}
Imputationswert	4.23	4.41	4.34	4.70	5.00	4.70	3.91	4.55	4.66

Einen weiteren Ansatz stellt die von *Dear (1959)* vorgeschlagene Ersetzung der fehlenden Ausprägungen durch den Gesamtmittelwert der Datenmatrix dar, der analog auf die Ersetzung fehlender Distanzen in einer Distanzmatrix übertragen werden kann. Dieses Verfahren führt jedoch im Fall einer Datenmatrix mit unterschiedlich skalierten Merkmalen zu Problemen und ist damit nur geeignet, wenn ausschließlich kardinale Merkmale mit vergleichbaren Ausprägungsmengen vorliegen.

4.2.1.2 Imputation des Verhältnisschätzers

Bei der Verwendung des Verhältnisschätzers zur Ersetzung der fehlenden Ausprägungen eines kardinalen Merkmals $k \in M$ der Datenmatrix A wird ein zusätzliches, ebenfalls kardinal skaliertes Merkmal $l \in M$, $l \neq k$ als sogenanntes Hilfsmerkmal benötigt, das mit dem Merkmal k hoch korreliert sein sollte. Mit $N_k = \{i: v_{ik} = 1\}$ kann der Verhältnisschätzer dann gemäß der Formel

$$a_{ik} = a_k^{Ratio} = \frac{1}{|N_l|} \cdot \frac{\sum_{j \in N_k} a_{jk}}{\sum_{j \in N_k} a_{jl}} \cdot \sum_{j \in N_l} a_{jl} \quad \forall\, i,k: v_{ik} = 0 \quad (k,l \in M, N_k \subset N_l) \qquad (4.18)$$

berechnet werden (vgl. *Ford, 1976, S. 324, Platek, Gray, 1983, S. 287-288*). Beim Merkmal l müssen für mindestens die Objekte, bei denen bezüglich des Merkmals k Ausprägungen vorliegen, ebenfalls Daten vorhanden sein, wobei im Fall $N_k = N_l$ der Verhältnisschätzer nach (4.18) dem Schätzer durch das arithmetische Mittel nach (4.14) entspricht. Für eine sinnvolle Verhältnisschätzung sollte daher zumindest die Bedingung $N_k \subsetneq N_l$ erfüllt sein, wobei $N_l = N$ den Idealfall darstellt.

Obwohl unter den Voraussetzungen, daß die fehlenden Daten des Merkmals k in einer linearen Beziehung zu den Ausprägungen des Hilfsmerkmals stehen und von den restlichen (m–2) Merkmalen der Datenmatrix unabhängig sind, die Bedingung MAR für die Anwendung des Verhältnisschätzers theoretisch ausreicht, muß aufgrund einer fehlenden Validierung dieser Voraussetzungen durch dieses Ersetzungsverfahren von der Annahme MCAR für die Daten ausgegangen werden (vgl. *Ford, 1976, S. 325*).

Beispiel:

In der Datenmatrix des Anhangs A (Fall 1) sind bei den drei kardinalen Merkmalen Deskriptive Statistik, Spezialgebiete und Statistische Grafiken fehlende Ausprägungen vorhanden. Unter Verwendung der in Abschnitt 4.1.1.2 berechneten Korrelationsmatrizen kann festgestellt werden, daß das Merkmal Multivariate Verfahren zu den Merkmalen Deskriptive Statistik und Spezialgebiete sowie das Merkmal Businessgrafiken zum Merkmal Statistische Grafiken jeweils die höchste Korrelation aufweist. Da die Ausprägungen der Merkmale Multivariate Verfahren und Businessgrafiken vollständig vorliegen, sind die Vorausset-

zungen der Formel (4.18) in jedem Fall erfüllt, so daß sich die folgenden Imputationswerte für die fehlenden Ausprägungen ergeben:

Fehlende Ausprägung	$a_{24}, a_{12,4}, a_{14,4}$	$a_{12,7}, a_{14,7}, a_{15,7}$	$a_{43}, a_{11,9}$
Verwendetes Hilfsmerkmal	Multivariate Verfahren	Multivariate Verfahren	Businessgrafiken
Imputationswert	70.14	45.11	39.09

4.2.1.3 Imputation mittels Zufallsauswahl

Im Rahmen einer Imputation mittels Zufallsauswahl sind die folgenden drei Ansätze zu unterscheiden, die in der Literatur zwar lediglich im Zusammenhang mit einer Datenmatrix erwähnt werden (vgl. z.B. *Schnell, 1986, S. 95*), aber teilweise auch auf eine Distanzmatrix übertragbar sind:

- Ersetzung fehlender Ausprägungen oder Distanzen durch eine aus einem Zufallszahlengenerator gezogene Zahl
- Ersetzung fehlender Ausprägungen eines Merkmals durch Zufallsauswahl aus den vorhandenen Ausprägungen bzw. Ersetzung fehlender Distanzen zweier Objekte durch Zufallsauswahl aus den vorhandenen Distanzen dieser Objekte
- Ersetzung aller fehlenden Merkmalsausprägungen eines Objekts durch die Ausprägungen eines zufällig ausgewählten vollständigen Objekts

Im Fall einer Ersetzung fehlender Ausprägungen oder Distanzen durch eine aus einem Zufallszahlengenerator gezogene Zahl stellt sich das Problem, welche Verteilung die Merkmalsausprägungen bzw. Distanzen jeweils besitzen und somit zur Bestimmung der Zufallszahlen zugrundegelegt wird. Diese Verteilung kann entweder a priori bekannt sein oder muß anhand der vorhandenen Werte angenommen bzw. bestimmt werden, wie dies beispielsweise von *Sonquist und Dunkelberg (1977, S. 270-271)* vorgeschlagen wird. Entsprechend können statistische Tests, wie der χ^2-Anpassungstest oder der Test von Kolmogoroff-Smirnoff, zur Überprüfung einer Verteilungsannahme herangezogen werden. Eine Beschreibung dieser Tests kann beispielsweise *Bamberg und Baur (1992, S. 198-202)* oder *Büning und Trenkler (1978, S. 85)* entnommen werden. Eine Darstellung der Möglichkeiten zur Bestimmung von Pseudozufallszahlen beliebiger Verteilungen findet sich z.B. bei *Neumann (1977, S. 289-295, 306-349)*. Für die nominal polytomen Merkmale einer Datenmatrix ergibt sich das Problem, daß die Verteilung der Merkmalsausprägungen lediglich in Form einer Wahrscheinlichkeitsfunktion angegeben werden kann und eine explizite Bestimmung der Verteilungsfunktion aufgrund der fehlenden Rangfolge der Merkmalsausprägungen nicht angemessen ist.

Beispiel:

Für die als Zufallsvariablen betrachteten Ausprägungen der Merkmale Preisniveau (1), Benutzeroberfläche (2), Programmierbarkeit (3), Deskriptive Statistik (4), Spezialgebiete (7) und Statistische Grafiken (9) der Datenmatrix des Anhangs A (Fall 1) werden die nachfolgend angegebenen Verteilungen bzw. Wahrscheinlichkeitsfunktionen unterstellt, wobei auf eine statistische Überprüfung im Rahmen dieses Beispiels verzichtet wird:

- Alle a_{i1} besitzen die Wahrscheinlichkeitsfunktion $f(a_{i1}) = \begin{cases} \frac{2}{13} & \text{für } a_{i1} = \text{niedrig} \\ \frac{5}{13} & \text{für } a_{i1} = \text{mittel} \\ \frac{2}{13} & \text{für } a_{i1} = \text{gehoben} \\ \frac{4}{13} & \text{für } a_{i1} = \text{hoch} \\ 0 & \text{sonst} \end{cases}$.
- Alle a_{i2} besitzen die Wahrscheinlichkeitsfunktion $f(a_{i2}) = \begin{cases} \frac{2}{14} & \text{für } a_{i2} = \text{M} \\ \frac{1}{14} & \text{für } a_{i2} = \text{K} \\ \frac{11}{14} & \text{für } a_{i2} = \text{M/K} \\ 0 & \text{sonst} \end{cases}$.
- Alle a_{i3} sind binomialverteilt mit $n = 1$ und $p = \frac{8}{13}$.
- Alle a_{i4} sind in $[0,100]$ gestutzt normalverteilt mit $\mu = 76.61$ und $\sigma = 13.79$.
- Alle a_{i7} sind in $[0,100]$ gestutzt normalverteilt mit $\mu = 44.47$ und $\sigma = 29.69$.
- Alle a_{i9} sind in $[0,100]$ gestutzt normalverteilt mit $\mu = 40.15$ und $\sigma = 25.84$.

Damit ergeben sich für die fehlenden Ausprägungen unter Verwendung eines Pseudozufallszahlengenerators die folgenden Imputationswerte:

Fehlende Ausprägung	a_{21}	$a_{10,1}$	a_{24}	a_{43}	a_{93}
Imputationswert	mittel	hoch	M/K	ja	ja

Fehlende Ausprägung	a_{24}	$a_{12,4}$	$a_{14,4}$	$a_{12,7}$	$a_{14,7}$	$a_{15,7}$	a_{49}	$a_{11,9}$
Imputationswert	66.60	76.96	70.10	38.65	53.46	35.88	3.56	41.88

Für die fehlenden paarweisen Distanzen der Distanzmatrix des Anhangs B ergeben sich unter Annahme einer Gleichverteilung der Distanzen im Intervall $[0,10]$ und unter Verwendung eines Pseudozufallszahlengenerators die folgenden Imputationswerte:

Fehlende Distanz	d_{14}	$d_{1,10}$	d_{27}	$d_{2,10}$	d_{35}	d_{37}	d_{46}	d_{59}	d_{69}
Imputationswert	6.11	7.88	5.73	3.47	4.76	2.59	2.26	7.26	7.22

Bei einer Ersetzung der fehlenden Ausprägungen eines Merkmals durch Zufallsauswahl aus den vorhandenen Ausprägungen bzw. einer Ersetzung der fehlenden Distanzen zweier Objekte durch Zufallsauswahl aus den vorhandenen Distanzen dieser Objekte stellt sich die Frage, ob den vorhandenen Werten jeweils die gleiche oder eine unter-

schiedliche Wahrscheinlichkeit einer Auswahl zugeordnet wird. In jedem Fall erfolgt implizit die Unterstellung einer bestimmten Verteilung der vorhandenen Merkmalsausprägungen bzw. Distanzen. Die Imputation von vorhandenen Ausprägungen bzw. Distanzen besitzt einen reinen Dopplungscharakter, d.h. nur bereits realisierte Werte der unvollständigen Daten- bzw. Distanzmatrix kommen als Schätzungen in Betracht.

Beispiel:

Wird den jeweils vorhandenen Merkmalsausprägungen bzw. Distanzen die gleiche Wahrscheinlichkeit einer Auswahl zugeordnet, ergeben sich für die fehlenden Ausprägungen der Merkmale Deskriptive Statistik (4), Spezialgebiete (7) und Statistische Grafiken (9) der Datenmatrix des Anhangs A (Fall 1) sowie für die fehlenden paarweisen Distanzen der Distanzmatrix des Anhangs B (Fall 1) die nachfolgend angegebenen Imputationswerte. Dabei werden die jeweils vorhandenen Werte zunächst numeriert. Entsprechend dieser Numerierung sind dann gleichverteilte, ganzzahlig gerundete Pseudozufallszahlen zu bestimmen, mittels denen auf die zu verwendenden Imputationswerte geschlossen werden kann. Auf eine Darstellung der Imputationswerte für die fehlenden Ausprägungen der Merkmale Preisniveau, Benutzeroberfläche und Programmierbarkeit wird in diesem Beispiel aufgrund der im Prinzip gleichen Verteilungsannahme wie bei der Ersetzung durch Zufallszahlen verzichtet.

Fehlende Ausprägung	a_{24}	$a_{12,4}$	$a_{14,4}$	$a_{12,7}$	$a_{14,7}$	$a_{15,7}$	a_{49}	$a_{11,9}$
Imputationswert	62.33	89.33	61.66	31.00	52.00	31.00	20.00	16.00

Fehlende Distanz	d_{14}	$d_{1,10}$	d_{27}	$d_{2,10}$	d_{35}	d_{37}	d_{46}	d_{59}	d_{69}
Imputationswert	4.32	2.57	3.45	5.33	4.96	4.45	4.45	3.02	5.65

Bei einer Ersetzung aller fehlenden Merkmalsausprägungen eines Objekts $i \in N$ durch die entsprechenden Ausprägungen eines zufällig ausgewählten vollständigen Objekts $j \in N, j \neq i$, der sogenannten **random imputation**, erfolgt für jeden unvollständigen Objektvektor a^i nur eine Zufallsauswahl zur Bestimmung des Objektvektors a^j (vgl. *Santos, 1981, S. 22*). Im Fall einer einzigen fehlenden Merkmalsausprägung bei Objekt i ist dieser Ansatz identisch zur oben dargestellten zufälligen Auswahl aus den vorhandenen Merkmalsausprägungen. Eine Übertragung der random imputation auf eine unvollständige Distanzmatrix ist aufgrund der Struktur dieser Matrix nicht möglich.

4.2.1.4 Imputation auf Basis von Expertenratings

Bei der Ersetzung durch Expertenratings wird ein fehlender Wert unter Verwendung der vorliegenden Informationen der Daten- oder Distanzmatrix durch einen Experten geschätzt (vgl. *Schnell, 1986, S. 96*). Dieses Verfahren entspricht damit im Grundgedanken dem der Regressionsanalyse, nämlich der Spezifizierung eines funktionalen Zu-

sammenhangs, wobei das Regressionsmodell sowie die Regressionskoeffizienten im Prinzip durch den Experten subjektiv festgelegt werden. Das von *Lösel und Wüstendörfer (1974, S. 350)* vorgeschlagene Heranziehen zusätzlicher, noch nicht vorliegender Informationen wäre zwar wünschenswert, scheint aber im Rahmen einer datenanalytischen Untersuchung weniger geeignet, da eine erneute Erhebung der noch fehlenden Daten der Erhebung weiterer Daten vorzuziehen ist.

Aufgrund des damit verbundenen Aufwands sowie der unter Umständen unzutreffenden Zusammenhangs- oder Kausalannahmen scheinen Expertenratings im Rahmen der Datenanalyse eine weniger geeignete Technik zur Schätzung fehlender Daten zu sein, so daß auf formale Prozeduren zurückgegriffen werden muß.

4.2.2 Imputation innerhalb von Klassen

Die Idee einer Imputation innerhalb von Klassen, man spricht in diesem Zusammenhang auch von sogenannten **Imputationsklassen**, liegt darin begründet, daß im allgemeinen innerhalb der Objektmenge einzelne Gruppen von Objekten existieren, wobei die Objekte derselben Gruppe sehr ähnlich und die unterschiedlicher Gruppen sehr verschieden sind. Diese Informationen können zur Bestimmung von Imputationswerten für die fehlenden Daten verwendet werden. Darüber hinaus benötigen die entsprechenden Imputationsverfahren lediglich die im Vergleich zur Eigenschaft MCAR weniger restriktive MCARC-Annahme, daß innerhalb der Imputationsklassen das Fehlen einzelner Werte in keiner Beziehung zu den Variablen der Untersuchung steht. Die im folgenden vorzustellenden Verfahren werden lediglich im Zusammenhang mit einer unvollständigen Datenmatrix betrachtet. Eine Übertragung auf den Fall einer unvollständigen, unmittelbar erhobenen Distanzmatrix ist dabei nicht möglich, da im Fall einer Distanzmatrix ein unsystematischer Ausfallmechanismus innerhalb von Klassen nicht definiert werden kann.

Die in Abschnitt 4.2.1 dargestellten einfachen Imputationstechniken können ausnahmslos auf das Konzept der Imputationsklassen übertragen werden. Gemäß dieser Überlegung erfolgt dann eine Imputation des **Klassenlageparameters** oder des **Klassenverhältnisschätzers** sowie durch **Zufallsauswahl innerhalb der Klassen** oder durch **Expertenratings auf Basis der Imputationsklassen**. Auf eine ausführliche Darstellung dieser Ansätze, die beispielsweise bei *Schnell (1986, S. 106-107)* zu finden ist, wird an dieser Stelle verzichtet. Statt dessen sollen zwei ausschließlich auf Imputationsklassen basierende MD-Verfahren, die sogenannten Cold-Deck- und Hot-Deck-Techniken, vorgestellt werden. Zunächst werden jedoch die Möglichkeiten der Bestimmung von Imputationsklassen diskutiert.

4.2.2.1 Bestimmung der Imputationsklassen

Neben der Zielsetzung möglichst homogener Klassen ist grundsätzlich zu beachten, daß in allen Imputationsklassen eine ausreichende Anzahl von vorhandenen Ausprägungen bei allen Merkmalen mit fehlenden Daten vorliegt. Ausreichend bedeutet in diesem Zusammenhang, daß mit dem jeweiligen MD-Verfahren die entsprechenden Imputationswerte zumindest noch bestimmt werden können. Die Forderung nach möglichst vielen vorhandenen Ausprägungen bei allen Merkmalen mit fehlenden Komponenten in den einzelnen Klassen ist nicht zweckmäßig, da sich dadurch ein Konflikt mit der Zielsetzung möglichst homogener Klassen ergibt.

In der Literatur (vgl. z.B. *Chapman, 1976, 1983a*) existieren eine Reihe von Vorschlägen zur Bestimmung der Imputationsklassen.[4] Die einfachste Möglichkeit besteht darin, ein vollständig vorliegendes, kategorisches Merkmal zur Klasseneinteilung zu verwenden. Natürlich können auch die klassierten Daten eines quantitativen Merkmals, ein externer Faktor oder mehrere Merkmale gleichzeitig zur Aufspaltung der Objektmenge herangezogen werden (vgl. z.B. *Rizvi, 1983, S. 306*). Da die auf den gebildeten Klassen basierenden Imputationsverfahren die Eigenschaft MCARC voraussetzen, muß mittels der in Kapitel 3 vorgestellten Ansätze zur Strukturanalyse einer unvollständigen Datenmatrix untersucht werden, ob die MCAR-Annahme innerhalb der Klassen gerechtfertigt ist. Diese Vorgehensweise, d.h. Festlegung der Imputationsklassen und anschließende Untersuchung der Eigenschaft MCARC, ist im allgemeinen jedoch nicht zweckmäßig, da für den Fall, daß die MCARC-Annahme nicht aufrechterhalten werden kann, eine andere Klassifikation bestimmt werden muß. Im Vergleich dazu besser geeignet erscheint die Vorgehensweise, zunächst den der gesamten Datenmatrix zugrundeliegenden Ausfallmechanismus zu untersuchen und anschließend auf Basis dieser Ergebnisse die Imputationsklassen zu bestimmen. In Abhängigkeit vom zugrundeliegenden Ausfallmechanismus innerhalb der gesamten Datenmatrix sind dabei die folgenden Ansatzpunkte, die im Anschluß noch ausführlich dargestellt werden, zu nennen:

- **Die Daten sind MCAR:** Damit ist die Eigenschaft MCARC in jedem Fall erfüllt. Eine Klassifikation der Objekte kann dann auf Basis beliebiger Merkmale unter Berücksichtigung der grundsätzlichen Zielsetzung, daß die Objekte innerhalb der Klassen möglichst ähnlich sind und im Hinblick auf das Ersetzungsverfahren ausreichend vorhandene Ausprägungen aufweisen, erfolgen.

---

[4] Die vorgeschlagenen Klassifikationsansätze werden zwar meist im Zusammenhang mit der Bestimmung von sogenannten Gewichtungsklassen im Rahmen der Anwendung von Gewichtungsmethoden erwähnt, können jedoch auch zur Festlegung der Imputationsklassen herangezogen werden (vgl. *Rizvi, 1983, S. 305*). Die Gewichtungsmethoden zählen grundsätzlich zu den Parameterschätzverfahren und werden daher erst in Abschnitt 4.3 betrachtet.

- **Die Daten sind MAR:** Unter Berücksichtigung des in Abschnitt 2.2.3 dargestellten Theorems sowie der dabei diskutierten Anwendbarkeit in realen datenanalytischen Untersuchungen ist eine Klassifikation der Objekte auf Basis aller vorhandenen Daten in der Art durchzuführen, daß die Innerklassendistanzen minimal werden. Dadurch wird die Wahrscheinlichkeit für das Vorliegen der Eigenschaft MCARC maximal.
- **Die Daten sind nicht MAR:** Die MCARC-Annahme nach (2.13) ist genau dann erfüllt, wenn die Wahrscheinlichkeiten für das Vorhandensein bzw. Fehlen der Merkmalsausprägungen aller Objekte der einzelnen Klassen identisch sind (vgl. Abschnitt 2.2.3 sowie *Oh, Scheuren, 1983, S. 148*, *Schnell, 1986, S. 98*). Die Forderung nach identischen Antwort- bzw. Ausfallwahrscheinlichkeiten innerhalb der Klassen ist jedoch nicht zweckmäßig, da im allgemeinen entweder die Anzahl der Objekte in den einzelnen Klassen oder die Anzahl der jeweils vorhandenen Merkmalsausprägungen für die Anwendung eines entsprechenden MD-Verfahrens zu gering sein wird. Selbst unter der zusätzlichen Forderung, daß die Ausfallwahrscheinlichkeiten kleiner als Eins sind, kann eine ausreichende Anzahl vorhandener Merkmalsausprägungen nicht garantiert werden. Gemäß diesen Überlegungen sind die Imputationsklassen unter der Zielsetzung möglichst homogener Antwort- bzw. Ausfallwahrscheinlichkeiten zu bestimmen, um somit näherungsweise das Vorliegen der Eigenschaft MCARC zu gewährleisten.

Wenn die Daten der Eigenschaft MCAR oder MAR genügen, können natürlich auch andere, nicht auf Imputationsklassen basierende MD-Verfahren zur Anwendung kommen. Eine Imputation innerhalb von Klassen ist unter Umständen jedoch durchaus zweckmäßig, da durch die Klassenbildung zusätzliche Informationen bereitgestellt werden.

Liegt die Eigenschaft MCAR für die Daten vor, dann kann die Bestimmung der Imputationsklassen auf Basis beliebiger Merkmale erfolgen. Dabei ist die Zielsetzung möglichst homogener Klassen und ausreichend vorhandener Ausprägungen in den Klassen zu berücksichtigen. In diesem Fall bieten sich die bereits vorgestellten Ansätze zur Berücksichtigung fehlender Daten an. So können lediglich die vollständig vorliegenden Merkmale für eine Klassifikation herangezogen werden. Des weiteren ist auch eine Vervollständigung der Datenmatrix mittels der in Abschnitt 4.2.1 vorgestellten einfachen Imputationstechniken denkbar. Dadurch kann eine Klassifikation auf Basis der gesamten Datenmatrix erfolgen. Darüber hinaus sind zur Bestimmung einer Distanzmatrix aus der unvollständigen Datenmatrix, die für einige Clusteranalyseverfahren benötigt wird, die in Abschnitt 4.1.2.2 dargestellten Möglichkeiten einer Verwendung der für die Objekte paarweise vorhandenen Merkmalsausprägungen geeignet. Da in allen Fällen letztendlich von einer vollständigen Datenbasis ausgegangen wird, eignen sich somit alle bekannten Clusteranalyseverfahren zur Bildung der Imputationsklassen.

Beispiel:

Für die Datenmatrix des Anhangs A (Fall 1) sollen die Imputationsklassen zunächst auf Basis der vollständig vorliegenden Merkmale Testverfahren, Multivariate Verfahren und Businessgrafiken gebildet werden. Verwendet man beispielsweise das Complete Linkage Verfahren, wobei zur Distanzbestimmung die City-Block-Metrik mit den merkmalsspezifischen Gewichten 1/Spannweite herangezogen wurde, dann ergibt sich die folgende 5-Klassen-Lösung, die gegenüber dem CLUDIA-Verfahren austauschinvariant ist:

Klasse 1	**Klasse 2**	**Klasse 3**	**Klasse 4**	**Klasse 5**
CRUNCH MICROSTAT II MINITAB RS/1 STATISTIX STATPAC GOLD	CSS NCSS SAS SPSS SYSTAT	STATA STATGRAPHICS	BMDP	P-STAT

Ausgehend von einer Distanzmatrix, die auf Basis der für die Objekte paarweise vorhandenen Merkmalsausprägungen bei Verwendung aller Merkmale der Datenmatrix bestimmt wurde, kommt man zur folgenden 6-Klassen-Lösung:

Klasse 1	**Klasse 2**	**Klasse 3**	**Klasse 4**	**Klasse 5**	**Klasse 6**
CRUNCH MICROSTAT II STATISTIX STATPAC GOLD	CSS STATA SYSTAT	BMDP P-STAT SPSS	SAS STATGRAPHICS	MINITAB RS/1	NCSS

Dabei wurde zur Aggregation der merkmalsweisen Distanzen, die jeweils auf das Intervall $[0;1]$ normiert wurden, die Formel (4.9) herangezogen. Die Klassifikation erfolgte wiederum mit dem Complete Linkage Verfahren, wobei sich die 6-Klassen-Lösung, für die das Ellenbogenkriterium spricht, bei Anwendung des CLUDIA-Verfahrens als austauschinvariant erweist.

Bezogen auf die gesamte Datenmatrix liegen in beiden Lösungen die Objekte der einelementigen Klassen vollständig vor und in den anderen Klassen ist bezüglich der Merkmale mit fehlenden Daten zumindest eine Ausprägung vorhanden. Damit ist beispielsweise eine Imputation des Klassenlageparameters oder die Anwendung der Cold- bzw. Hot-Deck-Techniken, die später noch vorgestellt werden, möglich.

Liegt lediglich die Eigenschaft MAR für die gesamte Datenmatrix vor, dann können die in Abschnitt 2.2.3 dargestellten Überlegungen bezüglich des Zusammenhangs zwischen den Eigenschaften MAR und MCARC zur Bestimmung der Imputationsklassen herangezogen werden. Danach wird die Wahrscheinlichkeit für das Vorliegen der Eigenschaft MCARC maximal, falls die Daten MAR sind und eine Klassifikation der Objekte auf Basis aller vorhandenen Daten mit dem Ziel der Minimierung der Innerklassendistanzen erfolgt. Diese „Approximation" erscheint für reale datenanalytische Untersuchungen

im Hinblick auf die damit verbundene Reduzierung des Arbeitsaufwands vertretbar. Die Bestimmung der Imputationsklassen kann dann in der Art erfolgen, daß zunächst eine Distanzmatrix auf Basis der für die Objekte paarweise vorhandenen Merkmalsausprägungen bestimmt wird. Dadurch wird eine Distanzbildung auf Basis aller vorhandenen Daten gewährleistet. Anschließend kann in Abhängigkeit der Klassenanzahl eine Minimierung der Innerklassendistanzen mit Hilfe des CLUDIA-Verfahrens erfolgen. Im Hinblick auf die Wahl einer geeigneten Klassenanzahl ist zu beachten, daß zum einen mit der Erhöhung der Klassenanzahl die Summe der Innerklassendistanzen monoton fällt. Zum anderen wird in den einzelnen Klassen eine für das jeweilige Imputationsverfahren ausreichende Anzahl von vorhandenen Ausprägungen bei allen Merkmalen mit fehlenden Komponenten benötigt.

Beispiel:

Für die Datenmatrix des Anhangs A wird der Fall 4 betrachtet, bei dem das Fehlen der Daten beim Merkmal Preisniveau von den Ausprägungen des Merkmals Deskriptive Statistik abhängt, d.h. die Daten MAR aber nicht OAR sind. Ausgehend von einer Distanzmatrix auf Basis der für die Objekte paarweise vorhandenen Ausprägungen bezüglich aller Merkmale ergeben sich die folgenden drei Klassen:

Klasse 1	**Klasse 2**	**Klasse 3**
BMDP	CRUNCH	CSS
P-STAT	MICROSTAT II	NCSS
RS/1	MINITAB	STATGRAPHICS
SAS	STATISTIX	SYSTAT
SPSS	STATPAC GOLD	
STATA		

Die Aggregation der merkmalsweisen Distanzen, die jeweils auf das Intervall $[0;1]$ normiert wurden, erfolgte mittels (4.9). Das gemäß dem Ellenbogenkriterium resultierende Ergebnis des Complete Linkage Verfahrens wurde dabei durch die Anwendung des CLUDIA-Verfahrens im Hinblick auf die Summe der Innerklassendistanzen verbessert. Eine feinere Klasseneinteilung und somit eine weitere Reduzierung der Summe der Innerklassendistanzen ist jedoch nicht möglich, da in diesem Fall die Objekte MINITAB, RS/1 und STATA eine eigene Klasse bilden. In dieser Klasse fehlen dann jedoch sämtliche Ausprägung des Merkmals Preisniveau.

Falls schließlich die Daten nicht MAR sind, dann können die Imputationsklassen unter der Zielsetzung möglichst homogener Antwort- bzw. Ausfallwahrscheinlichkeiten innerhalb der Klassen bestimmt werden. Die einfachste Möglichkeit besteht darin, die Objekte auf Basis ihrer MD-Muster zu klassifizieren. In diesem Fall stellt also die Indikatormatrix V den Ausgangspunkt einer Klassifikation dar. Da jedoch die Objekte im Hinblick auf die Ähnlichkeit ihrer MD-Muster klassifiziert werden, stellt sich in verstärktem Maße das Problem, daß unter Umständen alle Objekte einer Klasse keine vorhandenen

Ausprägungen bei den für die Imputation relevanten Merkmalen aufweisen. Folglich ist eine Bestimmung geeigneter Imputationsklassen auf Basis der Matrix V nicht immer möglich.

Beispiel:

Betrachtet man für die Datenmatrix des Anhangs A den Fall 2, bei dem ein systematischer Ausfallmechanismus vorliegt, dann führt eine Klassifikation auf Basis der Indikatormatrix V in jedem Fall zu einer 2-Klassen-Lösung, bei der alle Objekte mit fehlenden bzw. vorhandenen Ausprägungen bezüglich des Merkmals Multivariate Verfahren jeweils in einer Klasse sind. Damit sind zwar die Antwort- bzw. Ausfallwahrscheinlichkeiten innerhalb der Klassen sogar identisch, d.h. die Eigenschaft MCARC ist erfüllt, jedoch sind in einer Klasse keine Ausprägungen für das Merkmal Multivariate Verfahren vorhanden, so daß die Anwendung entsprechender Ersetzungstechniken nicht mehr möglich ist.

Ein Ausweg aus diesem Problem und damit eine andere Möglichkeit zur Bestimmung der Imputationsklassen gemäß dem Konzept homogener Antwort- bzw. Ausfallwahrscheinlichkeiten innerhalb der Klassen stellt die Anwendung des AID-Verfahrens dar (vgl. z.B. *Chapman, 1976, S. 248-249, Rizvi, 1983, S. 306*). Die diesem Ansatz zugrundeliegende Idee besteht darin, zunächst sukzessive die Merkmale $l \in M, l \neq k$ ($v_{\bullet l}^{ind} = 1$) zu bestimmen, mit denen die durch das MD-Muster des Merkmals k ($v_{\bullet k}^{ind} = 0$) vorgegebene Zerlegung der Objektmenge möglichst gut reproduziert werden kann. Da die auf diese Art ermittelten Merkmale das Vorhandensein bzw. Fehlen der Daten beim Merkmal k am besten erklären, können folglich die Imputationsklassen auf Basis dieser Merkmale festgelegt und zur Bestimmung der Imputationswerte für das Merkmal k herangezogen werden. Um das AID-Verfahren anwenden zu können, müssen die zur Erklärung des jeweiligen MD-Merkmalsmusters in Betracht zu ziehenden Merkmale nominal binäres Skalenniveau besitzen. Dabei liefert jedes dieser Merkmale ebenfalls eine Zerlegung der Objektmenge in zwei Klassen. In einem ersten Schritt wird nun zunächst das Merkmal ausgewählt, für dessen Zerlegung sich eine maximale Zwischengruppenvarianz bezüglich des abhängigen Merkmals, d.h. des jeweiligen MD-Merkmalsmusters ergibt. Anschließend wird für jede Klasse der im ersten Schritt gewählten Zerlegung eines der verbleibenden Merkmale bestimmt, für dessen Zerlegung wiederum die Zwischengruppenvarianz maximal wird. Diese Vorgehensweise wird solange fortgesetzt, bis entweder alle Merkmale zur Aufspaltung herangezogen wurden oder eine vollständige Erklärung des jeweiligen MD-Musters erreicht ist. Eine ausführliche Darstellung des AID-Verfahrens findet sich beispielsweise bei *Opitz (1980, S. 158-161)*.

Beispiel:

Betrachtet man für die Datenmatrix des Anhangs A wiederum den Fall 2, dann stellt die einzige zu erklärende Variable das MD-Muster des Merkmals Multivariate Verfahren dar. Jedoch kann aufgrund des Skalenniveaus auch lediglich das Merkmal Programmierbarkeit zur Identifikation nach dem AID-Verfah-

ren herangezogen werden. Mit der aus dem Merkmal Programmierbarkeit resultierenden Zerlegung der Objektmenge in zwei Klassen werden 33 Prozent der Gesamtvarianz der abhängigen Variable erklärt. Das Merkmal Programmierbarkeit impliziert schließlich die folgende 2-Klassen-Lösung:

Klasse 1	**Klasse 2**
BMDP	CRUNCH
CSS	MICROSTAT II
MINITAB	NCSS
P-STAT	STATGRAPHICS
RS/1	STATISTIX
SAS	STATPAC GOLD
SPSS	
STATA	
SYSTAT	

Für ein einziges, zur Identifikation herangezogenes Merkmal kann der Erklärungsanteil von 33 Prozent zwar als zufriedenstellend bezeichnet werden, jedoch sollten im allgemeinen weitere Merkmale zur Erklärung der Restvarianz ausgewählt werden können. Damit ist dieser Ansatz für die hier zugrundeliegende Datenmatrix nicht geeignet.

Die zur Anwendung des AID-Verfahrens notwendige Einschränkung, daß lediglich nominal binäre Merkmale zur Erklärung des MD-Musters herangezogen werden können, ist für praktische Anwendungen sehr unbefriedigend. Eine Erweiterung auf kardinale und unter Umständen auch ordinale Merkmale ist jedoch dadurch möglich, daß umgekehrt untersucht wird, inwieweit eine Erklärung der in der Datenmatrix vorliegenden Merkmale $l \in M$, $l \neq k$ ($v_{\bullet l}^{ind} = 1$) durch das MD-Muster des Merkmals k ($v_{\bullet k}^{ind} = 0$) möglich ist. Durch eine entsprechende Anwendung des AID-Verfahrens können damit letztendlich die Merkmale bestimmt werden, die durch das MD-Muster am besten erklärt werden. Die Imputationsklassen werden dann auf Basis dieser Merkmale festgelegt. Darüber hinaus stellen die jeweiligen Erklärungsanteile sinnvolle merkmalsspezifische Gewichte dar, falls zur Anwendung eines Klassifikationsverfahrens zunächst eine Distanzmatrix berechnet werden muß.

Beispiel:

Für die im vorherigen Beispiel verwendete Datenbasis ergeben sich für das ordinale Merkmal Preisniveau, das nominal binäre Merkmal Programmierbarkeit sowie für die kardinalen Merkmale Deskriptive Statistik, Testverfahren, Spezialgebiete, Businessgrafiken und Statistische Grafiken die in der nachfolgenden Tabelle angegebenen Varianzen sowie die durch den MD-Indikator des Merkmals Multivariate Verfahren erklärten Anteile. Die Bestimmung der Imputationsklassen kann dann auf Basis der Merkmale mit den höchsten Erklärungsanteilen erfolgen. Verwendet man beispielsweise alle Merkmale mit einem Erklärungsanteil von mindestens 0.25, dann ergibt sich bei einer Normierung der merkmalsweisen Distanzen auf das Intervall $[0;1]$ sowie einer Anwendung des CLUDIA-Verfahrens dieselbe 2-Klassen-Lösung wie im

vorherigen Beispiel. Dabei ist anzumerken, daß die Bestimmung von mehr als zwei Klassen nicht sinnvoll ist, da lediglich ein Merkmal mit fehlenden Daten vorliegt und somit theoretisch nur zwei Klassen mit jeweils identischen Antwort- bzw. Ausfallwahrscheinlichkeiten existieren können.

Merkmal	Gesamtvarianz	Erklärte Varianz	Erklärungsanteil
Preisniveau	1.16	0.27	0.23
Programmierbarkeit	0.24	0.08	0.33
Deskriptive Statistik	149.61	48.02	0.32
Testverfahren	118.92	4.30	0.04
Spezialgebiete	656.99	168.50	0.26
Businessgrafiken	1601.93	219.10	0.14
Statistische Grafiken	594.32	101.30	0.17

Anstelle der Merkmale mit den höchsten Erklärungsanteilen können zur Distanzbildung auch alle Merkmale herangezogen werden. In diesem Fall werden bei der Aggregation der merkmalsweisen, auf das Intervall [0;1] normierten Distanzen die einzelnen Erklärungsanteile als merkmalsspezifische Gewichte verwendet. Das CLUDIA-Verfahrens führt auch in diesem Fall zu den im vorherigen Beispiel dargestellten zwei Klassen.

Neben dem AID-Verfahren existieren natürlich auch andere Identifikationsverfahren, mit denen die Merkmale, die ein vorliegendes MD-Muster möglichst gut erklären, bestimmt werden können. Auf eine explizite Darstellung derartiger Ansätze, wie z.B. die Anwendung der Diskriminanzanalyse, wird jedoch an dieser Stelle verzichtet.

4.2.2.2 Cold-Deck-Verfahren

Ausgehend von den ermittelten Imputationsklassen erfolgt bei den Cold-Deck-Verfahren grundsätzlich eine Ersetzung der fehlenden Daten durch Werte aus einer externen Quelle (vgl. z.B. *Little, Rubin, 1987, S. 60*). Die Imputationswerte selbst stehen somit in keinem Zusammenhang zu den Daten der aktuellen Untersuchung. Als externe Quellen können ähnliche Untersuchungen zu einem früheren Zeitpunkt oder aktuelle, zusätzliche Erhebungsergebnisse herangezogen werden. In jedem Fall müssen die Daten der externen Quelle den ermittelten Imputationsklassen zugeordnet werden, so daß anschließend eine entsprechende Imputation innerhalb der Klassen erfolgen kann. Als Imputationswerte können entweder zufällig ausgewählte Werte oder geeignete Lageparameter herangezogen werden (vgl. *Chapman, 1976, S. 245, Schnell, 1986, S.109*).

Verwendet man beispielsweise die Daten einer früheren Untersuchung zur gleichen Thematik als externe Quelle, dann sind zunächst die Objekte der früheren Untersuchung den auf Basis der aktuellen Daten gebildeten Imputationsklassen zuzuordnen. Dies kann dadurch erreicht werden, daß entweder die Merkmale, die zur Bestimmung der Imputationsklassen herangezogen werden, auch für eine Klassifikation der Objekte der früheren Untersuchung verwendet werden oder eine Zuordnung anhand der Klasseneigenschaften erfolgt (vgl. *Chapman, 1976, S. 245*). In beiden Fällen wird damit jedoch vorausgesetzt, daß der früheren und der aktuellen Untersuchung identische Merkmale zugrunde liegen.

Zusammenfassend stellen die Cold-Deck-Verfahren aufgrund der benötigten externen Informationen sowie der daran geknüpften Voraussetzungen eine im allgemeinen wenig geeignete Imputationsmethode dar. Die Anwendung der Cold-Deck-Verfahren in realen datenanalytischen Untersuchungen wird somit eher eine Ausnahme sein.

4.2.2.3 Hot-Deck-Verfahren

Im Gegensatz zu den oben dargestellten Cold-Deck-Verfahren wird bei den Hot-Deck-Verfahren zur Bestimmung von Imputationswerten ausschließlich das vorliegende Datenmaterial verwendet (vgl. z.B. *Chapman, 1976, S. 245, Schnell, 1985, S. 53*). Um darüber hinaus eine klare Abgrenzung von anderen Imputationsmethoden auf Basis des vorliegenden Datenmaterials herzustellen, sind die Hot-Deck-Verfahren grundsätzlich durch eine Verdopplung bereits vorhandener Daten charakterisiert, d.h. die fehlenden Daten werden durch vorhandene Daten ersetzt (vgl. z.B. *Ford, 1983, S. 186, Sande, 1983, S. 341*). Da die Ersetzung der fehlenden Daten eines Objekts durch die Ausprägungen eines ähnlichen Objekts sinnvoller erscheint als die Ersetzung durch die Ausprägungen eines beliebig ausgewählten Objekts, erfolgt dieser Verdopplungsprozeß innerhalb der gebildeten Imputationsklassen (*Ford, 1983, S. 186*). Gemäß diesen definierten Eigenschaften ergibt sich somit die folgende Vorgehensweise der Hot-Deck-Verfahren (vgl. *Ford, 1976, S. 325*):

1. Die Imputationsklassen werden festgelegt.
2. Innerhalb der Klassen wird für jede fehlende Merkmalsausprägung eine vorhandene Ausprägung bezüglich desselben Merkmals ausgewählt.
3. Die fehlenden Daten werden durch die jeweils ausgewählten Werte ersetzt.

Durch die unterschiedlichen Möglichkeiten der Auswahl von Imputationswerten aus den vorhandenen Daten sind auch unterschiedliche Varianten von Hot-Deck-Verfahren

denkbar. In der Literatur (vgl. z.B. *Fellegi, Holt, 1976, S. 25-27, Schnell, 1985, S. 53*) werden die folgenden Verfahrenskategorien unterschieden:

- **Sequentielle Hot-Deck-Verfahren**
- **Simultane Hot-Deck-Verfahren**

Zunächst sollen die sequentiellen Hot-Deck-Verfahren betrachtet werden. Dazu ist in der nachfolgenden Abbildung 4.1 das Grundprinzip dieser Verfahren in Form eines Struktogramms dargestellt.

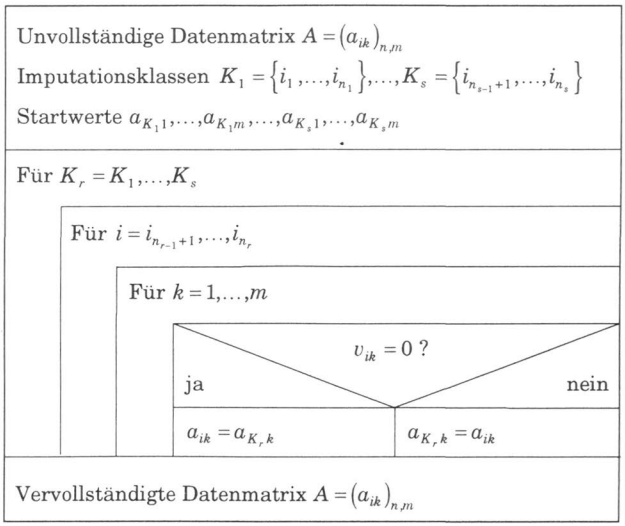

Abbildung 4.1: Sequentielle Hot-Deck-Verfahren

Ausgehend von einer unvollständigen Datenmatrix A werden in einem ersten Schritt innerhalb der festgelegten Imputationsklassen die Objekte in eine Reihenfolge gebracht und Startwerte für die zur Imputation heranzuziehenden Merkmalsausprägungen bestimmt. Abhängig von den verschiedenen Möglichkeiten, die Objektreihenfolge sowie die Startwerte festzulegen, ergeben sich unterschiedliche Varianten der sequentiellen Hot-Deck-Verfahren. Die Reihung der Objekte innerhalb der Imputationsklassen kann dabei zufällig, gemäß der Objektnummer oder nach der Ähnlichkeit der Objekte erfolgen (*Ford 1983, S. 196, Little, Rubin, 1987, S. 65*). Da jedoch eine Klassifikation der Objekte unter dem Gesichtspunkt möglichst homogener Klassen erfolgt, ist die zusätzliche Berücksichtigung der Ähnlichkeitsbeziehungen innerhalb der Imputationsklassen wenig zweckmäßig, zumal eine zufällige Reihenfolge der Objekte oder eine Reihung gemäß der

Objektnummer erheblich einfacher bestimmt werden können. Für die Festlegung der Startwerte gibt es in der Literatur (vgl. z.B. *Little, Rubin, 1987, S. 65, Schnell, 1986, S. 109*) ebenfalls mehrere Vorschläge. Danach können die Startwerte innerhalb der Klassen durch zufällige Auswahl einer vorhandenen Ausprägung für jedes Merkmal, durch die Klassenmittelwerte sowie durch die Anwendung eines Cold-Deck-Verfahrens ermittelt werden. Dabei erscheint die Verwendung der Klassenmittelwerte problematisch, da dadurch die Verdopplungseigenschaft der Hot-Deck-Verfahren verloren geht und somit keine klare Abgrenzung zur Mittelwertersetzung möglich ist. Sind schließlich die Reihenfolge der Objekte und die Startwerte festgelegt, dann erfolgt eine sequentielle Abarbeitung der gesamten Datenmatrix in der Art, daß innerhalb der Imputationsklassen geprüft wird, ob die Objekte der Reihe nach eine fehlende Ausprägung bezüglich der einzelnen Merkmale besitzen. Ist dies nicht der Fall, dann wird die entsprechende Ausprägung des Objekts zum neuen Imputationswert für das gerade betrachtete Merkmal innerhalb der Imputationsklasse, andernfalls wird dem Objekt der für das betrachtete Merkmal vorliegende, aktuelle Imputationswert der Klasse zugewiesen.

Beispiel:

Für die Datenmatrix des Anhangs A soll im folgenden der Fall 1 betrachtet werden. Dabei werden die in Kapitel 4.2.2.1 für diesen Fall beispielhaft auf Basis der vollständig vorliegenden Merkmale Testverfahren, Multivariate Verfahren und Businessgrafiken bestimmten Imputationsklassen herangezogen, die sich unter Verwendung der Objektnummern und einer Reihung der Objekte innerhalb der Klassen nach diesen Objektnummern wie folgt ergeben: $K_1 = \{2,4,5,8,13,14\}$, $K_2 = \{3,6,9,10,15\}$, $K_3 = \{11,12\}$, $K_4 = \{1\}$, $K_5 = \{7\}$. Da die Objekte der Klassen K_4 und K_5 keine fehlenden Daten aufweisen, müssen diese beiden Klassen nicht weiter berücksichtigt werden. Für die anderen Imputationsklassen sind nun zunächst die Startwerte festzulegen. Dazu erfolgt innerhalb der Klassen eine zufällige Auswahl aus den vorhandenen Ausprägungen der einzelnen Merkmale, wobei lediglich die Merkmale betrachtet werden müssen, die fehlende Daten aufweisen. In der nachfolgenden Arbeitstabelle sind beispielhaft für die Klasse K_1 die ausgewählten Startwerte sowie der sich anschließende, sequentielle Ersetzungsprozeß für die Merkmale Preisniveau (1), Benutzeroberfläche (2), Programmierbarkeit (3), Deskriptive Statistik (4), Spezialgebiete (7) und Statistische Grafiken (9) dargestellt:

	$a_{K_1 1}$	a_{i1}	$a_{K_1 2}$	a_{i2}	$a_{K_1 3}$	a_{i3}	$a_{K_1 4}$	a_{i4}	$a_{K_1 7}$	a_{i7}	$a_{K_1 9}$	a_{i9}
Startwert	mittel		M/K		ja		62.33		31.00		51.00	
$i = 2$		mittel	M/K		nein			62.33	0.00		11.25	
$i = 4$	niedrig			M/K		nein	61.66		2.66			11.25
$i = 5$	mittel		K		ja		62.33		31.00		28.24	
$i = 8$	hoch		M/K		ja		54.33		27.66		51.00	
$i = 13$	niedrig		M		nein		67.33		22.00		16.00	
$i = 14$	mittel		M/K		nein			67.33		22.00	21.25	

Anhand der Tabelle wird ersichtlich, daß bei den Objekten der Reihe nach den fehlenden Merkmalsausprägungen jeweils die aktuellen Imputationswerte der Klasse zugewiesen bzw. die vorhandenen Ausprägungen zu den neuen Imputationswerten der Klasse werden. Dabei wird ebenfalls deutlich, daß eine andere Startwertefestlegung oder Reihenfolge der Objekte auch zu einer anderen Imputation führen kann. Dieser Ersetzungsprozeß ist nun analog für die Klassen K_2 und K_3 durchzuführen, wobei auf eine Darstellung der Ergebnisse an dieser Stelle verzichtet wird.

Während bei den sequentiellen Hot-Deck-Verfahren jeweils merkmalsweise eine fehlende Ausprägung durch eine vorhandene ersetzt wird, erfolgt bei den simultanen Hot-Deck-Verfahren die Ersetzung sämtlicher fehlender Ausprägungen eines Objekts durch die Ausprägungen eines einzigen anderen Objekts. Das Grundprinzip der simultanen Hot-Deck-Verfahren ist in der Abbildung 4.2 in Form eines Struktogramms dargestellt.

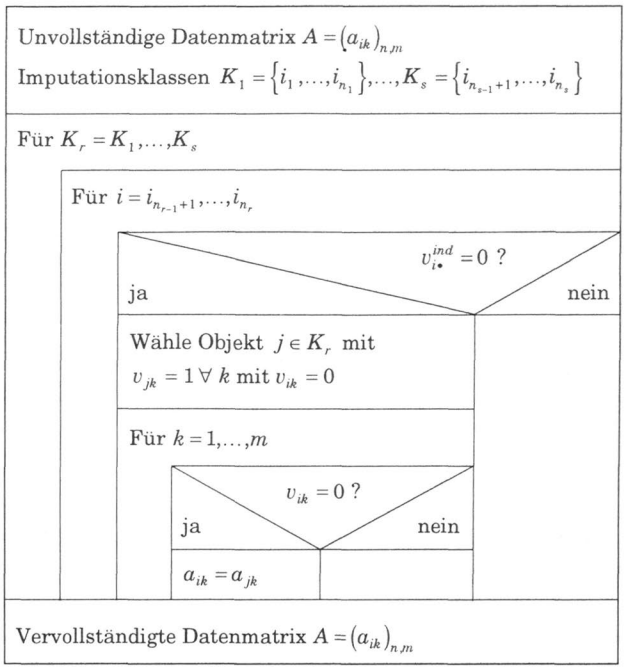

Abbildung 4.2: Simultane Hot-Deck-Verfahren

Innerhalb der Imputationsklassen kann das Objekt, das die Imputationswerte liefert, entweder zufällig oder bewußt im Sinne einer größtmöglichen Ähnlichkeit zum unvollständig vorliegenden Objekt bestimmt werden. Dadurch sind unterschiedliche Varianten dieses Verfahrens gekennzeichnet. In jedem Fall muß aber der Objektvektor, der die

Imputationswerte enthält, vorhandene Daten bezüglich aller Merkmale aufweisen, die im Rahmen der Ersetzung relevant sind. Um dies zu gewährleisten und darüber hinaus eine einfache Anwendung der Verfahren zu ermöglichen, wird in der Literatur meist die Verwendung der vollständig vorliegenden Objekte zur Festlegung der Imputationswerte vorgeschlagen. Der Begriff simultane Hot-Deck-Verfahren wird in Anlehnung an *Schnell (1986, S. 110)* verwendet, wobei auch die Bezeichnungen **joint imputation** (*Fellegi, Holt, 1976, S. 26*), **closest procedure** (*Ford, 1976, S. 326*), **nearest neighbor hot-deck** (*Little, Rubin, 1987, S.65*) und **random-choice procedure** (*Sande, 1983, S. 342*) unter diesem Begriff einzuordnen sind.

Beispiel:

Für die Datenmatrix des Anhangs A wird im folgenden wiederum der Fall 1 mit den Imputationsklassen $K_1 = \{2,4,5,8,13,14\}$, $K_2 = \{3,6,9,10,15\}$, $K_3 = \{11,12\}$, $K_4 = \{1\}$ und $K_5 = \{7\}$ betrachtet. Die Klassen K_4 und K_5 können aufgrund der vollständig vorliegenden Objekte unberücksichtigt bleiben. Innerhalb der Imputationsklassen sind dann für alle Objekte mit fehlenden Daten diejenigen Objekte zu bestimmen, die aufgrund ausreichend vorhandener Daten zur Imputation herangezogen werden können. Im einzelnen ergeben sich die folgenden Auswahlmöglichkeiten:

K_1		K_2		K_3	
$i = 2$	$j \in \{4,5,\mathbf{8},13\}$	$i = 9$	$j \in \{3,6,10,\mathbf{15}\}$	$i = 11$	$j \in \{\mathbf{12}\}$
$i = 4$	$j \in \{\mathbf{2},5,8,13,14\}$	$i = 10$	$j \in \{3,6,9,\mathbf{15}\}$	$i = 12$	$j \in \{\mathbf{11}\}$
$i = 14$	$j \in \{\mathbf{4},5,8,13\}$	$i = 15$	$j \in \{3,\mathbf{6},9,10\}$		

Die für den Fall einer zufälligen Auswahl jeweils zur Imputation heranzuziehenden Objekte sind in der Tabelle in Fettschrift hervorgehoben. Exemplarisch erhält man für die Klasse K_1 damit die folgenden Imputationen: a_{21} = hoch, a_{24} = 54.33, a_{42} = M/K, a_{43} = nein, a_{49} = 11.25, $a_{14,4}$ = 61.66, $a_{14,7}$ = 2.66. Analog ergeben sich die Imputationen für die Klassen K_2 und K_3, wobei auf eine explizite Darstellung an dieser Stelle verzichtet wird.

Aufgrund der Berücksichtigung der Ähnlichkeitsbeziehungen der Objekte durch die Verwendung von Imputationsklassen sowie der Verdopplungseigenschaft, besitzen die Hot-Deck-Verfahren zwei entscheidende Vorteile gegenüber den in Abschnitt 4.1 dargestellten einfachen Imputationstechniken. Zum einen kann, wie dies beispielsweise bei *Ford (1983, S. 187-189)* oder *Little und Rubin (1987, S. 63-64)* gezeigt wird, durch die Bestimmung möglichst homogener Imputationsklassen der Fehler bei der Schätzung der Merkmalsmittelwerte sowie der Stichprobenvarianzen reduziert werden. Unter der Annahme, daß die Verteilung der vorhandenen Daten und die Verteilung der fehlenden Werte identisch sind, ist aufgrund der Verdopplung vorhandener Daten zum anderen eine geringere Verzerrung der Verteilung der gesamten Stichprobe zu erwarten (*Ford, 1983, S. 190*).

Die Anwendung der Hot-Deck-Verfahren bringt jedoch auch einige Nachteile mit sich. So besteht die Gefahr, daß im Extremfall eine Ausprägung sehr oft dupliziert und somit die Verteilung der Merkmalsausprägungen erheblich verzerrt wird (*Ford, 1983, S. 196-197*). Des weiteren sind, wie beispielsweise *Schnell (1986, S. 110-111)* erwähnt, die sequentiellen Hot-Deck-Verfahren nicht invariant gegenüber der Reihenfolge der Objekte, so daß die Festlegung einer geeigneten Objektreihenfolge unter Umständen zu Problemen führt. Schließlich müssen innerhalb der Imputationsklassen ausreichend vorhandene Daten vorliegen. Im Fall der sequentiellen Hot-Deck-Verfahren bedeutet dies, daß innerhalb der Klassen für jedes Merkmal mit fehlenden Daten zumindest eine vorhandene Ausprägung vorliegen muß. Bei den simultanen Hot-Deck-Verfahren wird innerhalb der Klassen für alle zu vervollständigenden Objektvektoren jeweils ein Objekt benötigt, bei dem alle relevanten Ausprägungen vorliegen. Dies kann im einfachsten Fall durch einen vollständigen Objektvektor in jeder Imputationsklasse gewährleistet werden. Die dargestellten Minimalanforderungen an die in den Klassen vorhandenen Daten sind jedoch nur für die Funktionsfähigkeit der Algorithmen von Bedeutung. In praktischen Anwendungen sollten erheblich mehr Daten innerhalb der Imputationsklassen vorhanden sein. Das lediglich Vorliegen der minimal benötigten Daten führt zwangsläufig dazu, daß innerhalb der Klassen alle fehlenden Ausprägungen eines Merkmals durch denselben Wert ersetzt werden.

4.2.3 Multivariate Imputationstechniken

Bei Vorliegen einer unvollständigen Datenmatrix können die Zusammenhänge zwischen einzelnen Merkmalen aufgedeckt und für die Bestimmung der Imputationswerte genutzt werden. In diesem Fall stellen die Regressions-, die Varianz-, die Diskriminanz- sowie die Hauptkomponentenanalyse geeignete Methoden zur Schätzung der fehlenden Ausprägungen dar. Grundsätzlich muß zur Anwendung dieser Verfahren die Eigenschaft MCAR für die Daten vorliegen. Falls in den zugrundegelegten Modellen jedoch alle möglichen Abhängigkeitsbeziehungen zwischen den fehlenden Ausprägungen und den vorhandenen Daten adäquat berücksichtigt werden, reicht das Vorliegen der Eigenschaft MAR aus (vgl. *Little, Rubin, 1987, S. 45*).

Im Fall einer unvollständigen, unmittelbar erhobenen Distanzmatrix können zur Bestimmung von Imputationswerten die Eigenschaften von Distanzindizes herangezogen werden. Als Distanzeigenschaften werden die Dreiecksungleichung sowie die ultrametrische Ungleichung betrachtet. Mit diesen Eigenschaften können gewisse Abhängigkeiten zwischen den paarweisen Distanzen unterstellt bzw. gefordert und bei der Ermittlung von Schätzwerten für die fehlenden Distanzen berücksichtigt werden. Die entsprechenden Verfahren setzen die Eigenschaft MCAR für die paarweisen Distanzen voraus.

4.2.3.1 Imputation mittels Regressionsanalyse

Die Imputation fehlender Daten unter Verwendung von Regressionsmodellen zählt mit der Mittelwertersetzung zu den bekanntesten und in der Literatur am häufigsten erwähnten MD-Verfahren.[5] Grundsätzlich setzt die Regressionsanalyse kardinales Skalenniveau der Merkmale voraus, wobei für die unabhängigen Merkmale auch nominal binäres Skalenniveau zulässig ist. Im einfachsten Fall wird zwischen den Merkmalen ein linearer Zusammenhang unterstellt, so daß die **multiple lineare Regression** zur Anwendung kommen kann. Dieser Ansatz ist auch in den bislang veröffentlichten Arbeiten einer Imputation fehlender Daten mittels der Regressionsanalyse fast ausnahmslos vorzufinden und wird daher im folgenden weiter betrachtet. Die Idee der Verwendung **nichtlinearer Regressionsmodelle** ist in der Literatur kaum anzutreffen. Dies liegt vor allem an den damit verbundenen Problemen einer geeigneten Modellspezifikation und den auftretenden Schwierigkeiten bei der Schätzung der Modellparameter.

In Anlehnung an die Bezeichnung Zero-Order-Regression bei der Mittelwertersetzung wird in der Literatur der Begriff **First-Order-Regression** für die Ersetzung mittels der multiplen linearen Regression verwendet (vgl. z.B. *Toutenburg, 1992, S. 211, Afifi, Elashoff, 1966, S. 599*). Im Grundprinzip werden dabei auf Basis der aufgestellten Regressionsgleichungen und den aus den vorliegenden Ausprägungen gemäß dem Kleinst-Quadrate-Prinzip geschätzten Modellparameter schließlich Schätzwerte für die fehlenden Daten bestimmt. In Abhängigkeit der jeweils zu schätzenden fehlenden Ausprägung a_{ik} ($i \in N, k \in M$) ergeben sich die Regressionsgleichungen allgemein gemäß

$$\begin{pmatrix} a_{1k} \\ \vdots \\ a_{i-1,k} \\ a_{i+1,k} \\ \vdots \\ a_{nk} \end{pmatrix} = \begin{pmatrix} 1 & a_{11} & \cdots & a_{1,k-1} & a_{1,k+1} & \cdots & a_{1m} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & a_{i-1,1} & \cdots & a_{i-1,k-1} & a_{i-1,k+1} & \cdots & a_{i-1,m} \\ 1 & a_{i+1,1} & \cdots & a_{i+1,k-1} & a_{i+1,k+1} & \cdots & a_{i+1,m} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & a_{n1} & \cdots & a_{n,k-1} & a_{n,k+1} & \cdots & a_{nm} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{k-1} \\ \beta_{k+1} \\ \vdots \\ \beta_m \end{pmatrix}. \tag{4.19}$$

Die nach der Schätzung der Regressionskoeffizienten resultierende Gleichung zur Bestimmung des Imputationswerts besitzt dann die Form

$$a_{ik} = \beta_0 + \beta_1 a_{i1} + \ldots + \beta_{k-1} a_{i,k-1} + \beta_{k+1} a_{i,k+1} + \ldots + \beta_m a_{im} \quad (i \in N). \tag{4.20}$$

---

[5] Die Regressionsanalyse ist zwar im eigentlichen Sinn nicht multivariat, da der funktionale Zusammenhang zwischen einem abhängigen Merkmal und einer Reihe weiterer, unabhängiger Merkmale untersucht wird. In der Literatur (vgl. z.B. *Hartung, Elpelt, 1992* oder *Kendall, 1975*) erfolgt jedoch aufgrund der Bedeutung der Regressionsanalyse meist eine Einordnung unter die multivariaten Verfahren.

Eine ausführliche Darstellung der Parameterschätzung ist beispielsweise bei *Jobson (1991, S. 223-225)* oder *Bamberg und Schittko (1979, S. 17-19)* zu finden. Die in (4.19) und (4.20) angegebenen Gleichungen sind bezüglich der Dimensionen unter Umständen zu reduzieren, da aufgrund weiterer fehlender Daten nicht alle Objekte bzw. Merkmale berücksichtigt werden können oder das vorausgesetzte Skalenniveau bei einem Merkmal nicht vorliegt. Darüber hinaus müssen ohnehin nicht alle jeweils zu Verfügung stehenden Merkmale verwendet werden. So schlägt beispielsweise *Frane (1976, S. 410-411)* vor, lediglich das Merkmal mit der höchsten Korrelation zum abhängigen Merkmal als einziges unabhängiges Merkmal zu verwenden. In diesem Fall handelt es sich dann genau genommen um eine **einfache Regression**. Des weiteren können der Bestimmtheitskoeffizient sowie die Ergebnisse von Signifikanztests für die Regressionskoeffizienten herangezogen werden, um die unabhängigen Merkmale und damit das geeignete Regressionsmodell festzulegen.

Zur Bestimmung der in (4.19) allgemein dargestellten Regressionsgleichungen sind somit eine Reihe von Varianten denkbar (vgl. z.B. *Afifi, Elashoff, 1966, S. 599-600*). Diese Varianten unterscheiden sich hinsichtlich der jeweils verwendeten unabhängigen Merkmale sowie der zur Parameterschätzung herangezogenen Daten (vgl. *Berger, 1979, S. 395*). Im folgenden werden nun die in der Literatur bekanntesten Verfahren näher beschrieben.

Einen der ersten Ansätze stellt die **Methode von *Federspiel et al.*** *(1959)* dar.[6] Dabei handelt es sich um ein iteratives Regressionsverfahren, bei dem die fehlenden Komponenten bei einem Merkmal zunächst durch das arithmetische Mittel aus den vorhandenen Daten dieses Merkmals ersetzt werden. Danach wird für jedes ursprünglich fehlende Daten aufweisende Merkmal als abhängige Variable ein Regressionsmodell aufgestellt, wobei jeweils alle anderen Merkmale als unabhängige Variablen verwendet werden. Da alle n Objekte jeweils zur Schätzung der Modellparameter herangezogen werden, sind die in (4.19) dargestellten Regressionsgleichungen nach diesem Ansatz jeweils um eine Zeile zu ergänzen. Nach der Bestimmung der Regressionskoeffizienten erfolgt dann eine Schätzung sowie Imputation der fehlenden Ausprägungen mittels der Regressionsfunktionen. Mit den auf diese Art ersetzten fehlenden Daten wird der Prozeß der Modellbildung, Schätzung der Regressionskoeffizienten sowie Schätzung und Imputation der fehlenden Daten solange wiederholt, bis die Veränderungen der Imputationswerte eine vorgegebene Schranke unterschreiten (vgl. *Jackson, 1968, S. 837-839, Anderson et al., 1983, S. 458*). Die Konvergenz des Algorithmus von *Federspiel et al.* ist zwar noch nicht nachgewiesen, jedoch zeigt sich in den bislang durchgeführten Anwendungen dieser Technik, daß die Veränderungen der Imputationswerte bereits nach wenigen Ite-

[6] Zitiert nach *Jackson (1968, S. 837-839)* und *Anderson et al. (1983, S. 458)*.

rationen nur noch geringfügig sind (vgl. z.B. *Jackson, 1968, S. 839*). Um fundierte Aussagen über die Konvergenzeigenschaften der Methode von *Federspiel et al.* treffen zu können, soll zunächst ein Beispiel die Vorgehensweise dieses Ansatzes sowie einer im Prinzip ähnlichen, nicht iterativen Regressionsmethode illustrieren.

Beispiel:

Den Ausgangspunkt stellen die nachfolgend angegebenen Ausprägungen a_{ik} für zwei Merkmale k = 1,2 und fünf Objekte i = 1,2,3,4,5 dar, wobei die Ausprägung a_{52} nicht vorhanden ist:

i	1	2	3	4	5
a_{i1}	2	1	1	1	5
a_{i2}	10	8	6	0	

Gemäß der Methode von *Federspiel et al.* wird die fehlende Ausprägung zunächst durch den Merkmalsmittelwert ersetzt, d.h. a_{52} = 6. Anschließend wird mit dem Merkmal k = 1 als unabhängige und dem Merkmal k = 2 als abhängige Variable eine einfache Regression durchgeführt. Dabei ergeben sich die folgenden Schätzwerte für die Regressionskoeffizienten:

$$\hat{\beta}_1 = \frac{(2-2)\cdot(10-6)+(1-2)\cdot(8-6)+\ldots+(5-2)\cdot(6-6)}{(2-2)^2+(1-2)^2+\ldots+(5-2)^2} = \frac{1}{3},\ \hat{\beta}_0 = 6-\frac{4}{12}\cdot 2 = \frac{16}{3}.$$

Auf Basis der entsprechenden Regressionsfunktion erhält man den Imputationswert $a_{52} = \frac{16}{3}+\frac{1}{3}\cdot 5 = 7$.

Mit diesem Imputationswert kann eine erneute Schätzung der Regressionskoeffizienten sowie die anschließende Bestimmung eines neuen Imputationswerts erfolgen. Die im Laufe der Iterationen ermittelten Imputationswerte nähern sich dabei dem Wert 26 an. Setzt man a_{52} = 26, dann ergeben sich keine weiteren Veränderungen der Regressionskoeffizienten und damit auch des Imputationswerts.

Im folgenden wird nun ein nicht iterativer Ansatz betrachtet, der im Vergleich zur Methode von *Federspiel et al.* zur Schätzung der Modellparameter lediglich die Objekte verwendet, deren Ausprägungen beim abhängigen Merkmal vorhanden sind. In diesem Fall werden zur Bestimmung der Regressionskoeffizienten also die Objekte i = 1,2,3,4 herangezogen. Dabei erhält man die folgenden Schätzwerte:

$$\hat{\beta}_1 = \frac{(2-1.25)\cdot(10-6)+\ldots+(1-1.25)\cdot(0-6)}{(2-1.25)^2+\ldots+(1-1.25)^2} = \frac{16}{3},\ \hat{\beta}_0 = 6-\frac{16}{3}\cdot 1.25 = -\frac{2}{3},\ a_{52} = -\frac{2}{3}+\frac{16}{3}\cdot 5 = 26.$$

Der mit diesem nicht iterativen Ansatz ermittelte Imputationswert entspricht damit dem Imputationswert, der sich durch die Anwendung der Methode von *Federspiel et al.* als Grenzwert ergibt.

Die Erkenntnis, daß die Methode von *Federspiel et al.* und die im Beispiel dargestellte, nicht iterative Regressionsmethode zu den letztendlich gleichen Ergebnissen führen, kann zwar formal nicht bewiesen werden, ist aber intuitiv ersichtlich: Da die Imputationswerte beim abhängigen Merkmal immer so gewählt werden, daß sie der ermittelten Regressionsfunktion genügen, werden die Regressionsfunktion und damit die Imputati-

onswerte solange verändert, bis eine bestmögliche Anpassung an die fest vorgegebenen Daten erreicht ist. Zu den fest vorgegebenen Daten zählen neben den vorhandenen Ausprägungen auch die bei den unabhängigen Merkmalen gegebenenfalls fehlenden Werte, da diese durch den Merkmalsmittelwert ersetzt werden und im Lauf des Iterationsprozesses unverändert bleiben. Gemäß diesen Überlegungen scheint die Konvergenz der Methode von *Federspiel et al.* zwar gesichert, jedoch ist die Anwendung dieser Methode nicht zweckmäßig, da der im Beispiel dargestellte, nicht iterative Ansatz erheblich einfacher zu den gleichen Resultaten führt.[7]

Ein weiteres, in der Literatur häufig zitiertes Verfahren stellt die **Methode von *Buck*** *(1960)* dar. Dieses Verfahren ist dadurch gekennzeichnet, daß zum einen ausschließlich die vollständig vorliegenden Objekte zur Schätzung der Regressionskoeffizienten verwendet werden. Zum anderen wird für jede fehlende Ausprägung eines Objekts ein Regressionsmodell aufgestellt, das auf die vorhandenen Daten dieses Objekts zugeschnitten ist, d.h. als unabhängige Variablen werden alle Merkmale herangezogen, deren Ausprägungen bei diesem Objekt vorhanden sind. Damit ergibt sich im Fall ausschließlich quantitativer Daten die maximale Anzahl denkbarer Regressionsmodelle gemäß

$$m \cdot \sum_{i=1}^{m-1} \binom{m-1}{i} = m \cdot \left[\sum_{i=0}^{m-1} \binom{m-1}{i} - 1 \right] = m \cdot \left(2^{m-1} - 1\right). \tag{4.21}$$

Die Anzahl der tatsächlich benötigten Regressionsmodelle kann jedoch höchstens gleich der Anzahl der in der Datenmatrix insgesamt vorliegenden fehlenden Ausprägungen sein. Des weiteren wird die Anzahl der tatsächlich benötigten Regressionsmodelle durch das Vorliegen identischer MD-Objektmuster reduziert. Unter Verwendung des Ausdrucks im Zähler der Formel (3.30), der die Anzahl der in der Datenmatrix tatsächlich vorliegenden, unterschiedlichen MD-Objektmuster angibt, kann die Anzahl der Regressionsmodelle, die insgesamt aufzustellen sind, nach der Formel

$$v^{mis} - \sum_{i=1}^{n-1} v_{i\bullet}^{mis} \cdot \left[1 - \min \left\{ \sum_{j=i+1}^{n} \left(1 - v_{j\bullet}^{ind}\right) \cdot \left(1 - \min \left\{ \sum_{k=1}^{m} \left| v_{ik} - v_{jk} \right| ; 1 \right\} \right) ; 1 \right\} \right], \tag{4.22}$$

berechnet werden. Dabei wird von ausschließlich kardinal skalierten Merkmalen ausgegangen und die in der multiplen Regression grundsätzlich mögliche Verwendung nominal binärer Merkmale als unabhängige Variablen nicht berücksichtigt.

---

[7] Dabei wird vorausgesetzt, daß der Rang der Beobachtungsmatrix, d.h. die Matrix mit Einsen in der ersten Spalte sowie den Ausprägungen der unabhängigen Merkmale in den restlichen Spalten, gleich der Anzahl der unabhängigen Merkmale plus Eins ist. Nur unter dieser Annahme führen sowohl die Methode nach *Federspiel et al.* wie auch der im Beispiel dargestellte, nicht iterative Ansatz zu einer Lösung bei der Schätzung der Regressionskoeffizienten.

Beispiel:

Für die Datenmatrix des Anhangs A (Fall 1) werden im folgenden lediglich die sechs kardinal skalierten Merkmale betrachtet. Bei einer Anwendung der Methode von *Buck* sind gemäß der Formel (4.21) theoretisch $6 \cdot (2^{6-1} - 1) = 186$ unterschiedliche Regressionsmodelle denkbar, gemäß (4.22) jedoch insgesamt lediglich $8 - (0+0+0+1+0+0+0+0+0+0+0+2+0+0) = 5$ Regressionsmodelle aufzustellen. Im einzelnen ergeben sich die folgenden Regressionsfunktionen:

$$a_{i4} = -21.10 + 0.94a_{i5} + 0.80a_{i6} + 0.11a_{i7} - 0.37a_{i8} + 0.28a_{i9} \quad (i = 2),$$
$$a_{i4} = -27.03 + 1.02a_{i5} + 0.89a_{i6} - 0.38a_{i8} + 0.35a_{i9} \quad (i = 12,14),$$
$$a_{i7} = -87.88 + 1.06a_{i5} + 1.02a_{i6} - 0.30a_{i8} + 0.95a_{i9} \quad (i = 12,14),$$
$$a_{i7} = -88.73 + 0.39a_{i4} + 0.78a_{i5} + 0.72a_{i6} - 0.21a_{i8} + 0.89a_{i9} \quad (i = 15),$$
$$a_{i9} = 103.62 + 0.42a_{i4} - 1.43a_{i5} - 1.04a_{i6} + 0.39a_{i7} + 0.59a_{i8} \quad (i = 4,11).$$

Die Werte für das korrigierte Bestimmtheitsmaß liegen zwischen 0.76 und 0.89. Der nachfolgenden Tabelle können schließlich die resultierenden Imputationswerte für die fehlenden Ausprägungen entnommen werden:

Fehlende Ausprägung	a_{24}	$a_{12,4}$	$a_{14,4}$	$a_{12,7}$	$a_{14,7}$	$a_{15,7}$	a_{49}	$a_{11,9}$
Imputationswert	45.50	60.37	49.19	39.98	9.55	96.09	2.73	74.59

Im Vergleich zur Methode von *Federspiel et al.* besitzt die Methode von *Buck* die Vorteile der einfacheren Anwendbarkeit sowie der Verwendung ausschließlich vorhandener Daten zur Schätzung der Modellparameter und der fehlenden Ausprägungen. Im allgemeinen ist jedoch die Anzahl der aufzustellenden Regressionsmodelle erheblich höher und die Anzahl der vollständig vorliegenden Objekte kann unter Umständen für eine Schätzung der Modellparameter zu gering sein.[8]

Eine Modifikation des Ansatzes von *Buck* stellt die von ***Gleason und Staelin*** *(1975, S. 236-237)* vorgeschlagene Regressionsmethode dar. Bei dieser Methode werden die Regressionsmodelle gemäß dem Ansatz von *Buck* aufgestellt. Die Schätzung der Modellparameter basiert jedoch auf einer Datenmatrix, in der die fehlenden Ausprägungen bei den einzelnen Merkmalen jeweils durch den Merkmalsmittelwert der vorhandenen Daten ersetzt sind, d.h. zur Parameterschätzung werden alle n Objekte herangezogen. Auf Basis der ermittelten Regressionsfunktionen erfolgt anschließend die Bestimmung der Imputationswerte.

---

[8] Um die Kleinst-Quadrate-Schätzwerte der Regressionskoeffizienten durch Auflösung der Normalgleichungen mittels Matrixinversion berechnen zu können, muß die Anzahl der Beobachtungen größer als die Anzahl der unabhängigen Merkmale sein (vgl. z.B. *Bamberg, Schittko, 1979, S. 15*).

Ein weiteres Regressionsverfahren, das dem Ansatz von *Buck* sehr ähnlich ist und in der Literatur häufig zitiert wird, ist die **Methode von *Walsh*** *(1961)*. Im Vergleich zu *Buck*, der eine Schätzung der Regressionskoeffizienten auf Basis der ausschließlich vollständig vorliegenden Objekte vorschlägt, verwendet *Walsh* zur Schätzung der Parameter eines Regressionsmodells alle Objekte, die im Hinblick auf das jeweilige abhängige Merkmal eine vorhandene Ausprägung aufweisen. Die unter Umständen vorliegenden fehlenden Daten der unabhängigen Merkmale werden zuvor mittels Regressionsschätzungen ersetzt, um dadurch eine vollständige Datenbasis für die Parameterschätzung bereitzustellen. Da der Imputationsprozeß somit mehrstufig abläuft, ergibt sich die Notwendigkeit eines festgelegten Algorithmus, der in der Abbildung 4.3 in Form eines Struktogramms dargestellt ist. Dabei ist anzumerken, daß auf eine rein formale Darstellung verzichtet wurde, da die dazu notwendigen Ausführungen zu umfangreich sind (vgl. *Walsh, 1961, S. 205-208*). Statt dessen sind die einzelnen Schritte lediglich verbal beschrieben.

Der von Walsh vorgeschlagene Algorithmus beginnt damit, daß zunächst ein Parameter T festgelegt werden muß. Dieser Parameter steuert die Anzahl der zu berechnenden Regressionsfunktionen und damit die Geschwindigkeit, aber auch die Schätzgüte dieses Imputationsverfahrens. Abhängig von der Wahl des Parameters T sind der Reihe nach alle für die vorliegende Datenmatrix denkbaren Regressionsmodelle ohne unabhängige Variable, mit genau einer unabhängigen Variable, mit genau zwei unabhängigen Variablen usw. aufzustellen. Die Regressionskoeffizienten werden dabei aus den Daten aller Objekte geschätzt, die beim jeweiligen abhängigen Merkmal eine vorhandene Ausprägung aufweisen. Da die auf diese Art ausgewählten Objekte bei den unabhängigen Variablen jedoch unter Umständen fehlende Daten besitzen, muß die ausgewählte Datenbasis zur Parameterschätzung gegebenenfalls vervollständigt werden. Dies erfolgt unter Verwendung der in den vorhergehenden Schritten bereits bestimmten Regressionsfunktionen. Für $t = 1$ entfällt dieses Auffüllen der Daten und für $t = 3$ erfolgt für den Fall, daß bei einem Objekt lediglich eine Ausprägung bezüglich der beiden unabhängigen Merkmale fehlt, eine Imputation unter Verwendung der im Schritt 2 ermittelten, entsprechenden Regressionsfunktion auf Basis dieser beiden Merkmale und für den Fall, daß für ein Objekt die Ausprägungen bei beiden unabhängigen Merkmalen fehlen, eine Imputation durch die im Schritt 1 bestimmten Mittelwerte dieser Merkmale.

Wird der Parameter $T = m$ gewählt, dann sind $m \cdot 2^{m-1}$ Regressionsmodelle aufzustellen. Dies sind im Vergleich zu den bei der Methode von *Buck* denkbaren Modellen gemäß (4.21) m Modelle mehr, da *Buck* den Fall eines vollständig fehlenden Objektvektors und der damit verbundenen Ersetzung aller Merkmalsausprägungen durch das jeweilige arithmetische Mittel nicht betrachtet. Wird der Parameter $T < m$ gewählt, dann müssen zwar weniger Regressionsfunktionen berechnet werden, jedoch ist eine linearhomogene

Aggregation der im letzten Schritt bestimmten Regressionsfunktionen nötig. Damit ist die anschließende Ersetzung der fehlenden Ausprägungen bei einem Objekt unter Verwendung aller vorhandenen Daten möglich. Die Gewichtungen sind unter Verwendung der Schätzgüte der aggregierten Regressionsfunktionen festzulegen, wobei die Summe der Gewichtungen auf Eins normiert ist (vgl. *Walsh, 1961, S. 204, 207-208*).

Unvollständige quantitative Datenmatrix $A = (a_{ik})_{n,m}$, $T \leq m$

Für $t = 1, \ldots, T$

- Aufstellen von $m \cdot \binom{m-1}{t-1}$ Regressionsmodellen, wobei jedes der m Merkmale als abhängige Variable verwendet wird und jeweils alle Kombinationen $(t-1)$-ter Ordnung der restlichen $(m-1)$ Merkmale als unabhängige Variablen herangezogen werden.
- Schätzung der Modellparameter, wobei alle Beobachtungen mit einer vorhandenen Ausprägung bezüglich der jeweils abhängigen Variable herangezogen werden. Gegebenenfalls fehlende Daten bei den unabhängigen Variablen sind jeweils über eine geeignete, in den Schritten $t-1$, $t-2$, ..., 1 bereits ermittelte Regressionsfunktion mit diesen Variablen zu schätzen.

$T = m$?

ja

- Linearhomogene Aggregation aller für jeweils eines der m abhängigen Variablen in Schritt T bestimmten Regressionsfunktionen für alle m abhängigen Variablen als Bestimmungsgleichungen für die Ausprägungen der m Merkmale. Mit den Gewichtungen kann jeweils die Schätzgüte der einzelnen Regressionsfunktionen berücksichtigt werden.

nein

Ersetzung aller fehlenden Daten, wobei die Ersetzungswerte für die fehlenden Ausprägungen eines Objekts aus allen vorhandenen Ausprägungen dieses Objekts sowie unter Verwendung einer damit geeigneten Regressionsfunktion aus den Schritten 1 bis T bzw. der anschließend über linearhomogene Aggregation ermittelten Bestimmungsgleichung berechnet werden.

Vervollständigte Datenmatrix $A = (a_{ik})_{n,m}$

Abbildung 4.3: Methode von Walsh

An dieser Stelle ist noch anzumerken, daß in dem von *Walsh* angegebenen Algorithmus, analog zur Methode von *Buck*, nicht alle denkbaren Regressionsmodelle aufgestellt werden müssen. Die Betrachtung kann sich auf die Modelle beschränken, die für die vorliegende unvollständige Datenmatrix relevant sind.

Die Vorteile der Methode von *Walsh* liegen darin begründet, daß im Vergleich zur Methode von Buck im allgemeinen mehr Beobachtungen zur Schätzung der Modellparameter herangezogen werden. Des weiteren kann durch Parameter T der Rechenaufwand gesteuert werden, wobei gegebenenfalls Einbußen hinsichtlich der Schätzgüte in Kauf genommen werden müssen. Für $T = 1$ entspricht der Ansatz der Ersetzung durch das arithmetische Mittel. Als Nachteile sind die zusätzlich zu bestimmenden Regressionsfunktionen zur Ersetzung der fehlenden Daten der unabhängigen Merkmale sowie die dadurch resultierenden Verzerrungen bei der Bestimmung der Parameter anderer Regressionsmodelle zu nennen.

Eine auf den grundsätzlichen Ideen von *Buck* und *Walsh* basierende Variante einer Regressionsersetzung wird von ***Chan und Dunn*** *(1972, S. 474)* vorgeschlagen. Bei diesem Ansatz werden zunächst die fehlenden Daten aller Objekte mit lediglich einer fehlenden Ausprägung gemäß der Methode von *Buck* ersetzt. Anschließend werden unter Verwendung der bereits ermittelten Imputationswerte die fehlenden Daten aller Objekte mit genau zwei fehlenden Ausprägungen wiederum gemäß der Methode von *Buck* ersetzt. Dieser Prozeß wird solange durchgeführt, bis alle fehlenden Daten ersetzt sind und eine vollständige Datenmatrix resultiert. Durch die sofortige Ersetzung der fehlenden Daten wird also die Anzahl der dann vollständig vorliegenden Objekte laufend erhöht, so daß die Datenbasis, die zur Schätzung der Parameter weiterer Regressionsmodelle vorliegt, entsprechend vergrößert wird. Allerdings muß der bereits bei der Methode von *Walsh* aufgezeigte Nachteil in Kauf genommen, daß die Ergebnisse aufgrund der Verwendung von Regressionsschätzungen zur Bestimmung der Parameter anderer Regressionsmodelle verzerrt werden.

Weitere in der Literatur zu findende Varianten der bislang vorgestellten Regressionsmethoden sind durch die Verwendung der **schrittweisen Regression** (vgl. *Frane, 1976, S. 410-411*) sowie durch die **Verwendung von MD-Indikatorvariablen** als Regressoren (vgl. *Haitovsky, 1968, S. 68*) denkbar. Mit diesen beiden Ansätzen können unter Umständen die Abhängigkeitsbeziehungen der fehlenden Daten innerhalb der Datenmatrix besser berücksichtigt werden.

Wie bereits zu Beginn dieses Kapitels bemerkt wurde, gibt es eine Reihe von MD-Verfahren, die sowohl zur Bestimmung von Ersetzungswerten wie auch zur Parameterschätzung herangezogen werden können. Gerade die Ansätze im Bereich der Regressionsanalyse stellen eigentlich immer derartige Methoden dar, bei denen nach der Schät-

zung der Modellparameter auch Ersetzungswerte für die fehlenden Daten ermittelt werden können. In diesem Zusammenhang sind vor allem die **Methode von *Yates*** *(1933)* sowie ***Barlett's* ANCOVA Methode** *(1937)* zu nennen. Diese Verfahren sind im Hinblick auf deren grundsätzliche Intention, der Schätzung von Regressionskoeffizienten, zwar zu den Parameterschätzverfahren zu zählen, jedoch ist mit beiden Ansätzen eine Kleinst-Quadrate-Schätzung der fehlenden Daten nach der Bestimmung der Regressionskoeffizienten möglich (vgl. *Little, Rubin, 1987, S. 25-29, Toutenburg, 1992, S. 204-206*). Auf die Methode von *Yates* sowie auf *Barlett's* ANCOVA Methode wird im Rahmen der Parameterschätzverfahren in Abschnitt 4.3 noch einmal unter Berücksichtigung ihrer eigentlichen Zielsetzung eingegangen. Ein weiteres Parameterschätzverfahren, das auch Imputationswerte liefert, stellt der **EM-Algorithmus** (expectation maximization algorithm) von *Dempster et al. (1977)* dar. Nach der Festlegung von Startwerten für die zu schätzenden Parameter werden bei diesem Verfahren im ersten Schritt die Imputationswerte bestimmt. Dabei sind die vorhandenen Daten sowie die zu schätzenden Parameter heranzuziehen. Auf Basis der vervollständigten Datenmatrix erfolgt dann im zweiten Schritt eine Schätzung der Parameter nach dem Maximum-Likelihood-Prinzip. Diese beiden Schritte werden bis zur Konvergenz der geschätzten Parameter bzw. der Imputationswerte für die fehlenden Daten wiederholt (vgl. *Little, Rubin, 1987, S. 130*). Eine ausführliche Darstellung des EM-Algorithmus erfolgt in Abschnitt 4.3.2.

4.2.3.2 Imputation mittels Varianzanalyse

Mit der Varianzanalyse kann der Zusammenhang zwischen einem quantitativen Merkmal als abhängige Variable sowie einem oder mehreren nominalen Merkmalen als unabhängige Variablen untersucht werden. Damit stellt die Varianzanalyse im Prinzip einen Spezialfall der multiplen Regression dar, da sich durch eine geeignete Codierung der unabhängigen Merkmale die meisten Varianzanalysemodelle durch ein multiples Regressionsmodell ausdrücken lassen (*Jobson, 1991, S. 399*). Als derartige Codierungsmöglichkeit ist beispielsweise die sogenannte **Dummy-Codierung** zu nennen, bei der ein nominales Merkmale mit g Ausprägungen durch $(g - 1)$ nominal binäre Merkmale ersetzt wird. Da die Methoden der Varianzanalyse unter anderem zur Auswertung von Versuchsplänen herangezogen werden, existieren eine Vielzahl von unterschiedlichen Varianzanalysemodellen. Wird die Varianzanalyse zur Untersuchung von Abhängigkeitsbeziehungen bei einer gegebenen Datenmatrix und damit zur Bestimmung von Imputationswerten für die fehlenden Daten angewandt, dann ist zunächst aus der Vielzahl der möglichen Varianzanalysemodelle ein auf das vorliegende Datenmaterial zugeschnittenes Modell auszuwählen. Im folgenden wird beispielhaft lediglich das Modell einer zweifaktoriellen Varianzanalyse ohne Wechselwirkung der Faktoren sowie einer Be-

obachtung pro Zelle betrachtet. Weitere Varianzanalysemodelle sind unter anderem bei *Jobson (1991, S. 399-530)* oder *Hartung (1989, S. 607-636)* zu finden.

Das Grundmodell einer zweifaktoriellen Varianzanalyse ohne Wechselwirkung zwischen den Faktoren sowie einer Beobachtung pro Zelle, man spricht in diesem Zusammenhang auch von einem einfachen Blockexperiment, läßt sich folgendermaßen darstellen:

$$x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad \text{mit} \quad \sum_{i=1}^{r} \alpha_i = \sum_{j=1}^{s} \beta_j = 0 \quad (i = 1,\ldots,r,\ j = 1,\ldots,s). \tag{4.23}$$

Dabei bezeichnen x_{ij} den Beobachtungswert des abhängigen Merkmals bei der i-ten Ausprägung des ersten unabhängigen Merkmals (i-te Stufe des ersten Faktors) und der j-ten Ausprägung des zweiten unabhängigen Merkmals (j-te Stufe des zweiten Faktors), μ den Gesamtmittelwert, α_i bzw. β_j den Einfluß der i-ten Ausprägung des ersten bzw. der j-ten Ausprägung des zweiten unabhängigen Merkmals (Effekt der i-ten Stufe des ersten bzw. der j-ten Stufe des zweiten Faktors) und ε_{ij} die unabhängigen Fehlervariablen mit Mittelwert Null und Varianz σ^2. Die Kleinst-Quadrate-Schätzer für die Modellparameter sowie, daraus resultierend, für die Beobachtungswerte (vgl. z.B. *Jobson, 1991, S. 431*) ergeben sich gemäß

$$\hat{\mu} = \frac{1}{r \cdot s} \sum_{k=1}^{r} \sum_{l=1}^{s} x_{kl},\ \hat{\alpha}_i = \frac{1}{s} \sum_{l=1}^{s} x_{il} - \frac{1}{r \cdot s} \sum_{k=1}^{r} \sum_{l=1}^{s} x_{kl},\ \hat{\beta}_j = \frac{1}{r} \sum_{k=1}^{r} x_{kj} - \frac{1}{r \cdot s} \sum_{k=1}^{r} \sum_{l=1}^{s} x_{kl}, \tag{4.24}$$

$$\hat{x}_{ij} = \frac{1}{s} \sum_{l=1}^{s} x_{il} + \frac{1}{r} \sum_{k=1}^{r} x_{kj} - \frac{1}{r \cdot s} \sum_{k=1}^{r} \sum_{l=1}^{s} x_{kl}. \tag{4.25}$$

Dabei wird durch das Symbol ^ angedeutet, daß es sich jeweils um geschätzte Werte für den Gesamtmittelwert, die Effekte der beiden Faktoren und die Beobachtungswerte handelt.

Ein Ansatz, der das in (4.23) beispielhaft dargestellte Varianzanalysemodell zur Bestimmung von Imputationswerten für die fehlenden Daten heranzieht, stellt die **Methode von *Wilkinson*** *(1958)* dar. Dabei wird grundsätzlich vorausgesetzt, daß bezüglich der unabhängigen, nominalen Merkmale sämtliche Daten vorhanden sind und lediglich bezüglich des abhängigen, quantitativen Merkmals fehlende Ausprägungen vorliegen. Darüber hinaus ist aufgrund des gewählten Modells für jede Faktorstufenkombination maximal ein Beobachtungswert zulässig. Für jede Faktorstufenkombination, für die kein Beobachtungswert vorliegt, wird dann eine entsprechende Schätzgleichung gemäß (4.25) aufgestellt. Dabei werden jeweils alle zur Schätzung heranzuziehenden fehlenden Beobachtungswerte als Variablen betrachtet. Damit resultiert letztendlich ein Gleichungssystem mit genau soviel Unbekannten wie Faktorstufenkombinationen ohne

Beobachtungswerte (vgl. *Wilkinson, 1958, S. 264-268*). Die Variablen des Gleichungssystems können schließlich mittels der bekannten Lösungsverfahren, wie beispielsweise dem Gaußalgorithmus oder der Inversion der dann als regulär vorausgesetzten Koeffizientenmatrix, bestimmt werden (vgl. z.B. *Opitz, 1995, S. 247-248, 282, Wilkinson, 1958, S. 276-285*) und stellen die zur Imputation heranzuziehenden Werte dar.

Beispiel:

Für die Datenmatrix des Anhangs A (Fall 1) werden das Merkmal Deskriptive Statistik als abhängige Variable sowie die Merkmale Preisniveau und Programmierbarkeit als unabhängige Variablen betrachtet. Da gemäß dem oben formulierten Modell maximal ein Beobachtungswert pro Faktorstufenkombination vorliegen darf und sämtliche Daten bezüglich der unabhängigen Merkmale vorhanden sein müssen, werden im folgenden lediglich die Objekte NCSS, STATA, STATGRAPHICS, STATISTIX und SYSTAT betrachtet. Damit ist der Imputationswert für die fehlende Ausprägung des Objekts STATGRAPHICS zu bestimmen. Im einzelnen ergibt sich die folgende Tabelle, in der die den einzelnen Zellen zugrundeliegenden, vorhandenen und zu bestimmenden Beobachtungswerte sowie die entsprechenden Zeilen- bzw. Spaltensummen angegeben sind:

		Programmierbarkeit		**Zeilensumme**
		ja (1)	**nein (2)**	
Preisniveau	**gehoben (1)**	86.66	x_{12}	$86.66 + x_{12}$
	mittel (2)	65.66	76.66	142.32
	niedrig (3)	x_{31}	67.33	$67.33 + x_{31}$
Spaltensumme		$152.32 + x_{31}$	$143.99 + x_{12}$	$296.31 + x_{12} + x_{31}$

Ausgehend von diesen Werten erhält man die beiden folgenden Gleichungen zur Bestimmung der Imputationswerte für die fehlenden Daten:

$$\text{(I)} \quad x_{12} = \tfrac{1}{2}(86.66 + x_{12}) + \tfrac{1}{3}(143.99 + x_{12}) - \tfrac{1}{6}(296.31 + x_{12} + x_{31}),$$

$$\text{(II)} \quad x_{31} = \tfrac{1}{2}(67.33 + x_{31}) + \tfrac{1}{3}(152.32 + x_{31}) - \tfrac{1}{6}(296.31 + x_{12} + x_{31}).$$

Damit ergeben sich die Imputationswerte $x_{12} = 97.66$ und $x_{31} = 56.33$, wobei lediglich der Wert x_{12} zur Ersetzung der fehlenden Ausprägung des Objekts STATGRAPHICS benötigt wird. Dieselben Imputationswerte ergeben sich natürlich auch unter Verwendung einer Regressionsfunktion, die auf Basis der für dieses Beispiel vollständig vorliegenden Objekte NCSS, STATA, STATISTIX und SYSTAT mit einer entsprechenden Dummy-Codierung hinsichtlich der unabhängigen Merkmale bestimmt wird.

Wie anhand des Beispiels deutlich wird, stellt das Heranziehen eines vorher festgelegten, verhältnismäßig einfach handhabbaren Varianzanalysemodells keine geeignete Vorgehensweise dar. Jedoch wird die Formulierung eines auf die vorliegenden Daten zugeschnittenen Modells im allgemeinen ein unausgeglichenes Design, d.h. eine unterschiedliche Anzahl von Beobachtungswerten in den einzelnen Zellen zur Folge haben.

Somit ist die Bestimmung der Imputationswerte auf Basis eines Regressionsmodells mit entsprechender Codierung der unabhängigen Variablen meist zweckmäßiger ist als die Verwendung der Methode von *Wilkinson*.

Bei der von ***Rubin*** *(1972)* vorgeschlagenen Methode zur Ersetzung fehlender Daten mittels der Varianzanalyse wird ausgehend von dem gemäß (4.23) dargestellten Varianzanalysemodell vorausgesetzt, daß bezüglich der unabhängigen, nominalen Merkmale sämtliche Daten vorhanden sind, lediglich bezüglich des abhängigen, quantitativen Merkmals fehlende Ausprägungen vorliegen und für jede Faktorstufenkombination maximal ein Beobachtungswert gegeben ist. *Rubin* geht von der Überlegung aus, daß sich der Schätzwert für einen fehlenden Beobachtungswert aus der Linearkombination der für die fehlenden Daten gewählten Imputationswerte sowie der vorhandenen Daten ergibt, d.h. für den Fall eines einzigen fehlenden Beobachtungswerts x_{ij} aus der Gleichung

$$\hat{x}_{ij} = \alpha_{ij} x_{ij} + \sum_{k \neq i} \sum_{l \neq j} \alpha_{kl} x_{kl} \quad \left(i \in \{1,\ldots,r\}, j \in \{1,\ldots,s\} \right) \tag{4.26}$$

mit α_{ij} als entsprechende Gewichte resultiert (vgl. *Rubin, 1972, S. 138*). Um nun Kleinst-Quadrate-Schätzwerte für die Parameter des Varianzanalysemodells (4.23) zu erhalten, muß der fehlende Beobachtungswert x_{ij} so gewählt werden, daß das entsprechende Residuum $x_{ij} - \hat{x}_{ij}$ gleich Null wird, d.h.

$$x_{ij} - \hat{x}_{ij} = \left(1 - \alpha_{ij}\right) x_{ij} - \sum_{k \neq i} \sum_{l \neq j} \alpha_{kl} x_{kl} = 0 \quad \Leftrightarrow \quad x_{ij} = \frac{\sum_{k \neq i} \sum_{l \neq j} \alpha_{kl} x_{kl}}{\left(1 - \alpha_{ij}\right)}. \tag{4.27}$$

Den Ausdruck im Zähler auf der rechten Seite der in (4.27) dargestellten Äquivalenzbeziehung entspricht damit dem negativen Residuum, das sich für $x_{ij} = 0$ ergibt, während der Wert im Nenner das Residuum für $x_{ij} = 1$ sowie $x_{kl} = 0 \; \forall \; k \neq i, \; l \neq j$ darstellt. Diese beiden Residualwerte können schließlich mit den entsprechend modifizierten Daten unter Verwendung der in (4.25) angegebenen Schätzgleichung für die Beobachtungswerte bestimmt werden (vgl. *Rubin, 1972, S. 138*).

Für den Fall mehrerer fehlender Ausprägungen hinsichtlich des abhängigen Merkmals ergibt sich eine analoge Vorgehensweise. Eine entsprechende Darstellung, auf die an dieser Stelle verzichtet wird, ist der Arbeit von *Rubin (1972, S. 138-140)* zu entnehmen.

Wie sich durch die in (4.28), wiederum nur für den Fall eines einzigen fehlenden Beobachtungswerts angegebenen Umformungen zeigen läßt, ist die von *Rubin* vorgeschlagene Methode identisch zu dem Ansatz von *Wilkinson*. Den Ausgangspunkt stellt die gemäß der Methode von *Wilkinson* aufzustellende Gleichung dar, die sich unter Verwendung der Schätzgleichung für die Beobachtungswerte gemäß (4.25) sowie nach der ge-

mäß dem Kleinst-Quadrate-Prinzip geforderten Bedingung, daß der zu schätzende Beobachtungswert gleich dem entsprechenden, gesuchten Imputationswert sein soll, ergibt. Der am Ende resultierende Quotient enthält im Zähler das für $x_{ij} = 0$ sich ergebende, negative Residuum und im Nenner das Residuum, das für $x_{ij} = 1$ sowie $x_{kl} = 0 \; \forall \; k \neq i, \; l \neq j$ resultiert, und entspricht damit der von *Rubin* hergeleiteten Bestimmungsgleichung für den Imputationswert.

$$
\begin{aligned}
x_{ij} &= \frac{1}{s}\sum_{l=1}^{s} x_{il} + \frac{1}{r}\sum_{k=1}^{r} x_{kj} - \frac{1}{r \cdot s}\sum_{k=1}^{r}\sum_{l=1}^{s} x_{kl} \\
&= \frac{1}{s}\sum_{l \neq j} x_{il} + \frac{x_{ij}}{s} + \frac{1}{r}\sum_{k \neq i} x_{kj} + \frac{x_{ij}}{r} - \frac{1}{r \cdot s}\sum_{k \neq i}\sum_{l \neq j} x_{kl} - \frac{x_{ij}}{r \cdot s} \\
\Leftrightarrow \quad x_{ij}\left(1 - \left(\frac{1}{s} + \frac{1}{r} + \frac{1}{r \cdot s}\right)\right) &= \frac{1}{s}\sum_{l \neq j} x_{il} + \frac{1}{r}\sum_{k \neq i} x_{kj} - \frac{1}{r \cdot s}\sum_{k \neq i}\sum_{l \neq j} x_{kl} \qquad (4.28) \\
\Leftrightarrow \quad x_{ij} &= \frac{\frac{1}{s}\sum_{l \neq j} x_{il} + \frac{1}{r}\sum_{k \neq i} x_{kj} - \frac{1}{r \cdot s}\sum_{k \neq i}\sum_{l \neq j} x_{kl}}{1 - \left(\frac{1}{s} + \frac{1}{r} + \frac{1}{r \cdot s}\right)}.
\end{aligned}
$$

Ein weiterer, im Prinzip sehr einfacher Ansatz zur Bestimmung von Imputationswerten auf Basis eines Varianzanalysemodells ergibt sich aus der unmittelbaren Verwendung der Gleichung (4.25) zur Schätzung der fehlenden Daten. Dieser Ansatz basiert im Vergleich zu den bislang vorgestellten Methoden nicht auf dem Kleinst-Quadrate-Prinzip. Die Modellparameter werden statt dessen einfach aus den vorhandenen Daten bestimmt. Dabei sind zwei Varianten zu unterscheiden. Zum einen können lediglich die bezüglich des abhängigen Merkmals sowie die bezüglich der unabhängigen Merkmale vollständig vorliegenden Objekte herangezogen werden. Zum anderen können zusätzlich auch noch die Objekte verwendet werden, die zwar beim abhängigen Merkmal eine vorhandene Ausprägung, jedoch bei den unabhängigen Merkmalen fehlende Daten besitzen. In diesem Fall ist das Fehlen von Daten hinsichtlich der unabhängigen Merkmale durch eine entsprechende zusätzliche Kategorie zu berücksichtigen.

Beispiel:

Für die Datenmatrix des Anhangs A (Fall 1) werden wiederum das Merkmal Deskriptive Statistik als abhängige Variable sowie die Merkmale Preisniveau und Programmierbarkeit als unabhängige Variablen betrachtet. In der nachfolgenden Tabelle sind die jeweils in den einzelnen Zellen vorliegenden Beobachtungswerte sowie die entsprechenden Zeilen- bzw. Spaltenmittel zusammengefaßt. Die in Klammern angegebenen Ausprägungen stammen von Objekten, die fehlende Daten bezüglich der unabhängigen Merkmale aufweisen. Bei den ebenfalls in Klammern stehenden Mittelwerte werden diese Ausprägungen entsprechend berücksichtigt.

		Programmierbarkeit			Zeilenmittel
		ja (1)	nein (2)	fehlend (3)	
Preisniveau	hoch (1)	92.66, 92.66, 54.33		(89.66)	79.88 (78.33)
	gehoben (2)	86.66			86.66
	mittel (3)	80.33, 62.33, 65.66	76.66		71.25
	niedrig (4)		67.33	(61.66)	67.33 (64.50)
	fehlend (5)	(89.33)			(89.33)
Spaltenmittel		76.38 (78.00)	72.00	(75.66)	75.40 (76.61)

Die dann gemäß der Gleichung (4.25) resultierenden Schätzungen $\hat{x}_{22}$ = 83.26 (82.05) und $\hat{x}_{32}$ = 67.85 (66.64) sowie $\hat{x}_{52}$ = 72.00 (84.72) stellen damit die Imputationswerte für die fehlenden Ausprägungen der Objekte STATGRAPHICS, STATPAC GOLD und CRUNCH beim Merkmal Deskriptive Statistik dar. Die in Klammern angegebenen Schätzungen ergeben sich dabei aus den in der Tabelle in Klammern stehenden Werten. Beim Schätzwert $\hat{x}_{52}$ = 72.00 für das Objekt CRUNCH ist noch anzumerken, daß im Gegensatz zum Schätzwert $\hat{x}_{52}$ = (84.72) das Fehlen der Ausprägung beim Merkmal Preisniveau und damit der Einfluß dieses Faktors nicht berücksichtigt wurde.

4.2.3.3 Imputation mittels Diskriminanzanalyse

Mit der Diskriminanzanalyse soll, wie bei der Regressionsanalyse, der funktionale Zusammenhang zwischen einem abhängigen Merkmal und im allgemeinen mehreren unabhängigen Merkmalen aufgedeckt werden. Das abhängige Merkmal ist dabei jedoch nominal skaliert und für die unabhängigen Merkmale wird kardinales Skalenniveau vorausgesetzt. Damit kann die Diskriminanzanalyse zur Bestimmung von Imputationswerten für die fehlende Daten bei nominal skalierten Merkmalen herangezogen werden. Im folgenden sollen dieser Ansatz vorgestellt werden, wobei lediglich der Fall eines linearen funktionalen Zusammenhangs zwischen dem abhängigen und den unabhängigen Merkmalen betrachtet wird. Eine ausführliche Darstellung der linearen Diskriminanzanalyse ist beispielsweise bei *Opitz (1980, S. 146-151)* oder *Backhaus et al. (1990, S. 161-220)* zu finden.

Den Ausgangspunkt der Betrachtung stellt ein nominal skaliertes Merkmal $k \in M$ mit fehlenden Daten dar, wobei die s unterschiedlichen Merkmalsausprägungen eine Zerlegung der Objektmenge N in s Klassen implizieren. Die unabhängigen Merkmale $l \in M$, $l \neq k$ besitzen kardinales Skalenniveau. Zur Bestimmung des Imputationswerts für eine fehlende Ausprägung a_{ik} ($i \in N$, $k \in M$) unter Verwendung der linearen Diskriminanzanalyse werden zunächst die Gewichtungen $g_1,\ldots,g_{k-1},g_{k+1},\ldots,g_m$ der unabhängigen

Merkmale, die sogenannten **Diskriminanzkoeffizienten**, gesucht, so daß die gewichtete Summe

$$\begin{pmatrix} y_{1k} \\ \vdots \\ y_{i-1,k} \\ y_{i+1,k} \\ \vdots \\ y_{nk} \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1,k-1} & a_{1,k+1} & \cdots & a_{1m} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{i-1,1} & \cdots & a_{i-1,k-1} & a_{i-1,k+1} & \cdots & a_{i-1,m} \\ a_{i+1,1} & \cdots & a_{i+1,k-1} & a_{i+1,k+1} & \cdots & a_{i+1,m} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{n1} & \cdots & a_{n,k-1} & a_{n,k+1} & \cdots & a_{nm} \end{pmatrix} \begin{pmatrix} g_1 \\ \vdots \\ g_{k-1} \\ g_{k+1} \\ \vdots \\ g_m \end{pmatrix} \tag{4.29}$$

die vorgegebene Zerlegung möglichst gut reproduziert. Denkbare Ansätze zur Schätzung der Diskriminanzkoeffizienten im Hinblick auf das gewählte **Diskriminanzkriterium** sind beispielsweise bei *Bausch und Opitz (1993, S. 101-102)* dargestellt. Die maximale Anzahl der dabei resultierenden Gewichtungsvektoren ergibt sich bei $(m - 1)$ unabhängigen Merkmalen sowie s Ausprägungen des abhängigen Merkmals aus dem Minimum von $(m - 2)$ und s. Nach der Schätzung der Diskriminanzkoeffizienten kann nun für jeden ermittelten Gewichtungsvektor der **Diskriminanzwert** y_{ik} $(i \in N,\ k \in M)$ für die fehlende Ausprägung a_{ik} gemäß der **Diskriminanzfunktion**

$$y_{ik} = g_1 a_{i1} + \ldots + g_{k-1} a_{i,k-1} + g_{k+1} a_{i,k+1} + \ldots + g_m a_{im} \tag{4.30}$$

bestimmt werden. Mit diesen Werten ist schließlich eine Zuordnung des Objekts $i \in N$ zu einer der s Klassen und damit die Festlegung der entsprechenden Klassenausprägung bezüglich des abhängigen Merkmals als Imputationswert möglich. Die Zuordnung kann beispielsweise mittels der **Minimum-Distanz-Regel** (vgl. z.B. *Backhaus et al., 1990, S. 188-190)* oder im einfachsten Fall über einen Vergleich der Global- und Klassenmittel hinsichtlich der Diskriminanzwerte für jede Diskriminanzfunktion erfolgen.

Analog zu den Ausführungen bei den Imputationsverfahren mittels der Regressionsanalyse sind die in (4.29) und (4.30) angegebenen Gleichungen bezüglich der Dimensionen unter Umständen zu reduzieren, da aufgrund weiterer fehlender Daten nicht alle Objekte bzw. Merkmale berücksichtigt werden können oder das vorausgesetzte Skalenniveau bei einem Merkmal nicht vorliegt. Darüber hinaus müssen auch nicht alle jeweils zu Verfügung stehenden Merkmale verwendet werden. In Anlehnung an die im Rahmen der Imputation mittels der Regressionsanalyse beschriebenen Verfahren sind damit auch entsprechende Varianten des oben dargestellten Ansatzes denkbar. Diese Varianten unterscheiden sich vor allem hinsichtlich der in den Diskriminanzanalysemodellen verwendeten unabhängigen Merkmale sowie der zur Schätzung der Diskriminanzkoeffizienten herangezogenen Daten.

Beispiel:

Für die Datenmatrix des Anhangs A (Fall 1) soll im folgenden die fehlende Ausprägung des Objekts MICROSTAT II beim nominal polytomen Merkmal Benutzeroberfläche unter Verwendung der fünf kardinalen Merkmale Deskriptive Statistik, Testverfahren, Multivariate Verfahren, Spezialgebiete und Businessgrafiken geschätzt werden. Das kardinal skalierte Merkmal Statistische Grafiken kann aufgrund der dort ebenfalls fehlenden Ausprägung des Objekts MICROSTAT II als weiteres unabhängiges Merkmal nicht herangezogen werden. Die Diskriminanzanalyse wird auf Basis der Objekte durchgeführt, deren Ausprägungen bei allen verwendeten Merkmalen vollständig vorliegen. Als Diskriminanzkriterium wird eine Maximierung der Quotienten aus Zwischenklassenvarianz und Innerklassenvarianzsumme gewählt. Dabei ergeben sich die folgenden Diskriminanzfunktionen:

$$y_{i2} = 0.818a_{i4} - 0.209a_{i5} - 0.105a_{i6} - 0.442a_{i7} + 0.285a_{i8} \quad \text{(1. Diskriminanzfunktion)},$$
$$y_{i2} = 0.708a_{i4} - 0.564a_{i5} - 0.332a_{i6} - 0.242a_{i7} + 0.109a_{i8} \quad \text{(2. Diskriminanzfunktion)}.$$

Die erste Diskriminanzfunktion trennt bei einem Globalmittel der Diskriminanzwerte von 36.12 die Objekte mit der Ausprägung „M/K" (Klassenmittel der Diskriminanzwerte: 39.67) von den Objekten mit der Ausprägung „M" (Klassenmittel der Diskriminanzwerte: 30.55) bzw. „K" (Klassenmittel der Diskriminanzwerte: 22.35). Die zweite Diskriminanzfunktion trennt schließlich bei einem Globalmittel der Diskriminanzwerte von –6.63 die Objekte mit der Ausprägung „M" (Klassenmittel der Diskriminanzwerte: –8.78) von den Objekten mit der Ausprägung „K" (Klassenmittel der Diskriminanzwerte: –4.40). Die Anzahl der fehlerhaft klassifizierten Objekte beträgt 10 Prozent. Ausgehend von diesen Ergebnissen kann nun eine Zuordnung des Objekts MICROSTAT II zu einer der drei Klassen und damit die Bestimmung eines Imputationswerts für die fehlende Ausprägung beim Merkmal Benutzeroberfläche erfolgen. Mit den Diskriminanzwerten 31.25 (1. Diskriminanzfunktion) und –6.29 (2. Diskriminanzfunktion) wird das Objekt MICROSTAT II dabei der Klasse von Objekten mit der Ausprägung „K" zugeordnet. Diese Ausprägung ist somit als Imputationswert heranzuziehen.

Abschließend ist noch darauf hinzuweisen, daß als Imputationswerte im Fall einer Ersetzung mittels der Diskriminanzanalyse lediglich die Merkmalsausprägungen herangezogen werden, die in der unvollständigen Datenmatrix bereits vorliegen. Es können also nicht alle denkbaren Merkmalsausprägungen in Betracht gezogen werden.

4.2.3.4 Imputation mittels Hauptkomponentenmethode

Die Bestimmung von Imputationswerten mittels der Hauptkomponentenmethode setzt kardinales Skalenniveau für die Merkmale voraus. Da die Imputationswerte für die fehlenden Daten unter Verwendung einer Linearkombination aus den berechneten Faktorladungen sowie den Faktorwerten ermittelt werden, handelt es sich bei diesem Ansatz im Grunde um ein multiples Regressionsmodell (*Lösel, Wüstendörfer, 1974, S. 351*). In der Literatur existieren im wesentlichen zwei, nur geringfügig voneinander abweichende Ansätze einer Imputation mittels der Hauptkomponentenmethode, die im fol-

genden näher vorgestellt werden.[9] Eine ausführliche Darstellung der grundsätzlichen Vorgehensweise im Rahmen einer Hauptkomponentenanalyse kann beispielsweise *Jobson (1992, S.346-388)* entnommen werden.

Den ersten Ansatz einer Ersetzung mittels der Hauptkomponentenmethode stellt die **Methode von *Dear*** *(1959)* dar.[10] Die unvollständige quantitative Datenmatrix A wird dabei zunächst in eine standardisierte und vervollständigte Datenmatrix $\tilde{A}$ transformiert. Die Standardisierung erfolgt auf Basis der für jedes Merkmal vorhandenen Ausprägungen und alle fehlenden Daten werden durch den Wert Null, also durch den Mittelwert der standardisierten Daten, ersetzt, d.h.

$$\tilde{A}=(\tilde{a}_{ik})_{n,m} \text{ mit } \tilde{a}_{ik}=\begin{cases}\dfrac{a_{ik}-\bar{a}_k}{\sqrt{s_{kk}}} & \text{falls } v_{ik}=1\\ 0 & \text{falls } v_{ik}=0\end{cases}, \bar{a}_k=\frac{1}{|N_k|}\sum_{j\in N_k} a_{jk}, s_{kk}=\frac{1}{|N_k|}\sum_{j\in N_k}(a_{jk}-\bar{a}_k)^2. \quad (4.31)$$

Auf Basis dieser Matrix bzw. der daraus berechneten Korrelationsmatrix wird dann eine Hauptkomponentenanalyse durchgeführt. Als Ergebnis erhält man die **Ladungsmatrix** F mit den **Faktorladungen** als Koeffizienten sowie die **Faktorwertematrix** X mit den **Faktorwerten** als Koeffizienten und damit die Beziehung

$$\tilde{A}=X\cdot F^T=\begin{pmatrix} x_{11} & \cdots & x_{1m}\\ \vdots & & \vdots\\ x_{n1} & \cdots & x_{nm}\end{pmatrix}\begin{pmatrix} f_{11} & \cdots & f_{m1}\\ \vdots & & \vdots\\ f_{1m} & \cdots & f_{mm}\end{pmatrix}=\tilde{A}\cdot F\cdot F^T. \quad (4.32)$$

Die Ladungsmatrix ist orthogonal, d.h. $FF^T = F^TF = I$, und enthält spaltenweise die Eigenvektoren der aus der Matrix $\tilde{A}$ berechneten Korrelationsmatrix. Die Faktorwertematrix ergibt sich aus dem Produkt der vervollständigten Matrix $\tilde{A}$ und der Ladungsmatrix. *Dear* verwendet zur Bestimmung der Imputationswerte lediglich die erste **Hauptkomponente**, d.h. die erste Spalte der Faktorwertematrix X. Diese Hauptkomponente erklärt den größten Anteil der Ausgangsinformation. Eine Schätzung für die standardisierte Datenmatrix im Sinne des Kleinst-Quadrate-Kriteriums ergibt sich aufgrund der Approximationseigenschaft der Singulärwertzerlegung damit gemäß

$$\hat{\tilde{A}}=\begin{pmatrix} x_{11}\\ \vdots\\ x_{n1}\end{pmatrix}\begin{pmatrix} f_{11} & \cdots & f_{m1}\end{pmatrix}=\begin{pmatrix} \tilde{a}_{11} & \cdots & \tilde{a}_{1m}\\ \vdots & & \vdots\\ \tilde{a}_{n1} & \cdots & \tilde{a}_{nm}\end{pmatrix}\begin{pmatrix} f_{11}\\ \vdots\\ f_{m1}\end{pmatrix}\begin{pmatrix} f_{11} & \cdots & f_{m1}\end{pmatrix}. \quad (4.33)$$

---

[9] Ein Überblick der bekannten Ansätze einschließlich zweier Methoden, die aufgrund der geringen praktischen Bedeutung hier nicht betrachtet werden, findet sich bei *Stewart (1982, S. 397-398)*.

[10] Zitiert nach *Timm (1970, S. 419-421)* sowie *Chan und Dunn (1972, S. 474, 1974, S. 672)*.

Durch das Symbol ^ wird angedeutet, daß es sich um eine Schätzung für die standardisierte Datenmatrix handelt. Die geschätzte standardisierte Datenmatrix kann durch eine entsprechende Umkehrung der in (4.31) angegebenen Transformation wieder in eine nicht standardisierte Datenmatrix umgewandelt werden, die letztendlich die Imputationswerte für die fehlenden Daten enthält. Zusammenfassend ergibt sich gemäß der Formel

$$a_{ik} = \frac{1}{|N_k|}\sum_{j\in N_k} a_{jk} + \left[\frac{1}{|N_k|}\sum_{j\in N_k}\left(a_{jk} - \frac{1}{|N_k|}\sum_{j\in N_k} a_{jk}\right)^2\right]\cdot\left[f_{k1}\sum_{l=1}^{m}\tilde{a}_{il}f_{l1}\right] \quad \forall i,k: v_{ik} = 0 \qquad (4.34)$$

der Imputationswert für eine fehlende Ausprägung a_{ik}, wobei die benötigten Faktorladungskoeffizienten der berechneten Ladungsmatrix F zu entnehmen sind.

Beispiel:

Für die Datenmatrix des Anhangs A (Fall 1) werden im folgenden lediglich die sechs kardinal skalierten Merkmale Deskriptive Statistik (4), Testverfahren (5), Multivariate Verfahren (6), Spezialgebiete (7), Businessgrafiken (8) und Statistische Grafiken (9) betrachtet. In der nachfolgenden Tabelle sind für die einzelnen Merkmale die Mittelwerte und Standardabweichungen auf Basis der jeweils vorliegenden Daten sowie die Faktorladungen der ersten Hauptkomponente (Erklärungsanteil: 53.28 Prozent), die auf Basis der Korrelationsmatrix des gemäß (4.31) transformierten Datenmaterials berechnet wurden, dargestellt.

Merkmal k	4	5	6	7	8	9
Mittelwert $\bar{a}_k$	76.61	72.53	35.45	44.47	54.07	40.15
Standardabweichung $\sqrt{s_{kk}}$	13.20	10.90	20.96	28.42	40.02	24.82
Faktorladung f_{k1}	−0.3733	0.0153	−0.5088	−0.5026	−0.4650	−0.3644

Unter Anwendung der Formel (4.34) ergeben sich dann die folgenden Imputationswerte für die fehlenden Ausprägungen der betrachteten Merkmale, wobei die Berechnung beispielhaft für die fehlende Ausprägung a_{24} dargestellt ist:

Fehlende Ausprägung	a_{24}	$a_{12,4}$	$a_{14,4}$	$a_{12,7}$	$a_{14,7}$	$a_{15,7}$	a_{49}	$a_{11,9}$
Imputationswert	66.40	78.30	72.38	49.37	32.21	72.82	19.23	39.02

$$a_{24} = 76.61 + 13.20\cdot\left[-0.3733\cdot\left(0\cdot(-0.3733) + \frac{62.00-72.53}{10.90}\cdot 0.0153 + \ldots + \frac{11.25-40.15}{24.82}\cdot(-0.3644)\right)\right] = 66.40\,.$$

An dieser Stelle ist noch darauf hinzuweisen, daß die Ersetzung aller fehlenden Ausprägungen in der standardisierten Datenmatrix durch den Wert Null zwar zu keiner Veränderung der Mittelwerte für die einzelnen Merkmale, jedoch zu einer Reduzierung der durch die Standardisierung auf Eins normierten, merkmalsweisen Varianzen in der ver-

vollständigten Matrix führt. Alternativ können vor einer Standardisierung der Datenmatrix die fehlenden Ausprägungen durch die entsprechenden Mittelwerte der einzelnen Merkmale ersetzt werden. Da sich in beiden Fällen jedoch die gleiche Korrelationsmatrix und damit auch die gleiche Ladungsmatrix ergibt, sind die im allgemeinen geringfügigen Abweichungen der letztendlich resultierenden Imputationswerte lediglich auf die unterschiedliche Datenbasis für die Standardisierung zurückzuführen.

Zwei mögliche Erweiterungen der von *Dear* vorgeschlagenen Methode sind unmittelbar ersichtlich. Zum einen können mehr als nur eine Hauptkomponente zur Ersetzung der fehlenden Daten herangezogen werden. Zum anderen ist eine iterative Anwendung des Verfahrens in der Art vorstellbar, daß die Bestimmung der Imputationswerte nach (4.34) jeweils unter Verwendung der im vorhergehenden Schritt bereits nach (4.34) ermittelten Imputationswerte solange wiederholt wird, bis die Abweichungen zwischen den resultierenden Imputationswerten in zwei aufeinanderfolgenden Schritten unter einen vorgegebenen Wert fallen. In der von ***Gleason und Staelin*** *(1975)* vorgeschlagenen Imputationsmethode mittels der Hauptkomponentenanalyse werden diese beiden Erweiterungsmöglichkeiten berücksichtigt. Darüber hinaus wird der Ansatz von *Dear* geringfügig modifiziert. Den Ausgangspunkt der Methode von *Gleason und Staelin* stellt wiederum die gemäß (4.31) standardisierte und mit dem Wert Null für die fehlenden Ausprägungen vervollständigte Datenmatrix dar. Wie bei der Methode von *Dear* wird dann auf Basis dieser Matrix bzw. der daraus berechneten Korrelationsmatrix eine Hauptkomponentenanalyse durchgeführt. Die Bestimmung der Imputationswerte unter Verwendung der Faktorladungen erfolgt allerdings abweichend vom Ansatz von *Dear*. Dazu werden zunächst die Matrizen $\tilde{A}$ und F jeweils in zwei Teilmatrizen $\tilde{A}_{obs}$ und $\tilde{A}_{mis}$ bzw. F_{obs} und F_{mis} aufgespalten, d.h.

$$\tilde{A} = \left(\tilde{A}_{obs}, \tilde{A}_{mis}\right) = \left(\left(\tilde{a}_{ik}\right)_{n,q}, \left(\tilde{a}_{ik}\right)_{n,m-q}\right), \quad F = \begin{pmatrix} F_{obs} \\ F_{mis} \end{pmatrix} = \begin{pmatrix} \left(f_{kl}\right)_{q,m} \\ \left(f_{kl}\right)_{m-q,m} \end{pmatrix} \quad (q \geq 1). \tag{4.35}$$

$\tilde{A}_{obs}$ enthält die Ausprägungen $\tilde{a}_{ik}$ der q Merkmale ohne fehlende Daten und F_{obs} die Faktorladungen für diese Merkmale, während $\tilde{A}_{mis}$ die Teilmatrix der Ausprägungen $\tilde{a}_{ik}$ bezüglich der $(m - q)$ Merkmale mit fehlenden Daten und F_{mis} die Teilmatrix der entsprechenden Faktorladungen darstellen. Analog zu (4.32) gilt damit die Beziehung

$$\left(\tilde{A}_{obs}, \tilde{A}_{mis}\right) = \left(\tilde{A}_{obs}, \tilde{A}_{mis}\right) \cdot \begin{pmatrix} F_{obs} \\ F_{mis} \end{pmatrix} \cdot \left(F_{obs}^T, F_{mis}^T\right) = \left(\tilde{A}_{obs}, \tilde{A}_{mis}\right) \cdot \begin{pmatrix} F_{obs}F_{obs}^T & F_{obs}F_{mis}^T \\ F_{mis}F_{obs}^T & F_{mis}F_{mis}^T \end{pmatrix}. \tag{4.36}$$

Ausgehend von diesen Teilmatrizen soll nun die Matrix $\tilde{A}_{mis}$ mittels der Matrix $\tilde{A}_{obs}$ sowie der ersten t Spaltenvektoren der Ladungsmatrix F rekonstruiert werden (vgl.

Gleason, Staelin, 1975, S.233-234). Die Matrix dieser ersten t **Ladungsvektoren** wird mit $F^{(t)}$ bezeichnet und es gilt

$$F^{(t)} = \left(f_{kl}\right)_{m,t} = \begin{pmatrix} F_{obs}^{(t)} \\ F_{mis}^{(t)} \end{pmatrix} = \begin{pmatrix} \left(f_{kl}\right)_{q,t} \\ \left(f_{kl}\right)_{m-q,t} \end{pmatrix}. \tag{4.37}$$

Unter Verwendung von (4.36) und (4.37) erhält man die Schätzung $\hat{\tilde{A}}_{mis}$ für die Matrix $\tilde{A}_{mis}$ aus der Matrix $\tilde{A}_{obs}$ sowie der ersten t Ladungsvektoren wie folgt:

$$\hat{\tilde{A}}_{mis} = \tilde{A}_{obs} \cdot F_{obs}^{(t)} \cdot F_{mis}^{(t)^T} + \hat{\tilde{A}}_{mis} \cdot F_{mis}^{(t)} \cdot F_{mis}^{(t)^T} \Leftrightarrow \hat{\tilde{A}}_{mis} = \tilde{A}_{obs} \cdot F_{obs}^{(t)} \cdot F_{mis}^{(t)^T} \cdot \left(I - F_{mis}^{(t)} \cdot F_{mis}^{(t)^T} \right)^{-1}. \tag{4.38}$$

Dabei wird vorausgesetzt, daß die Matrix $(I - F_{mis}^{(t)} \cdot F_{mis}^{(t)^T})$ regulär ist. Damit ergibt sich eine Obergrenze für die maximal mögliche Anzahl der heranzuziehenden Ladungsvektoren, da eine Invertierung dieser Matrix nicht mehr möglich ist, sobald durch die Verwendung der ersten t Ladungsvektoren die Koeffizienten der restlichen Ladungsvektoren bezüglich aller Merkmale mit fehlenden Daten gleich dem Wert Null sind. Eine exakte Bestimmung der maximal möglichen Anzahl der verwendbaren Ladungsvektoren ist jedoch nicht möglich (vgl. *Gleason, Staelin, 1975, S. 235*). Grundsätzlich sollte die Auswahl der letztendlich herangezogenen Hauptkomponenten ohnehin im Hinblick auf deren Erklärungsanteil erfolgen, zumal die Ergebnisse im Fall einer Berücksichtigung von Hauptkomponenten mit einem geringen Erklärungsanteil sehr stark von der verwendeten Genauigkeit für die Faktorladungen abhängen.

Wie bei der Methode von *Dear* ist die Matrix $\hat{\tilde{A}}_{mis}$ schließlich wieder in eine nicht standardisierte Matrix zu transformieren, um somit die Imputationswerte für die fehlenden Daten zu erhalten. Darüber hinaus empfehlen *Gleason und Staelin* zur Verbesserung der Schätzungen für die Korrelationsmatrix und damit auch der Imputationswerte die bereits angesprochene iterative Anwendung des Verfahrens (vgl. *Gleason, Staelin, 1975, S. 238-239*).

Beispiel:

Für die Datenmatrix des Anhangs A (Fall 1) werden wiederum lediglich die sechs kardinal skalierten Merkmale betrachtet, wobei die Merkmale Testverfahren (5), Multivariate Verfahren (6) und Businessgrafiken (8) keine fehlenden Daten und die Merkmale Deskriptive Statistik (4), Spezialgebiete (7) und Statistische Grafiken (9) fehlende Daten aufweisen. Da die Berechnung der Ladungsmatrix analog zur Methode von *Dear* erfolgt, können die Mittelwerte und Standardabweichungen der Merkmale auf Basis der vorliegenden Daten sowie die berechneten Faktorladungen der ersten Hauptkomponente dem vorhergehenden Beispiel entnommen werden. In Ergänzung dazu sind nachfolgend noch die Faktorladungen der zweiten und dritten Hauptkomponente (Erklärungsanteile: 21.67 und 15.21 Prozent) angegeben. Auf die Darstellung weiterer Ladungsvektoren wird aus Gründen der geringen Erklärungsanteile verzichtet.

Merkmal k	4	5	6	7	8	9
Faktorladung f_{k2}	−0.4947	−0.6635	−0.0642	−0.0787	0.1059	0.5418
Faktorladung f_{k3}	−0.4396	0.6622	−0.3193	0.0691	0.4134	0.3010

Da die ersten beiden Hauptkomponenten bereits 74.95 Prozent der Ausgangsinformation erklären, sollen mit den entsprechenden Ladungsvektoren die Imputationswerte für die fehlenden Daten unter Anwendung von (4.38) bestimmt werden. Mit

$$\tilde{A}_{obs} = \begin{pmatrix} -1.5165 & 1.1369 & -1.3511 \\ -0.9661 & -0.9418 & -0.8513 \\ \vdots & \vdots & \vdots \\ -0.7826 & 1.2190 & 0.8728 \end{pmatrix}, \quad F_{obs}^{(2)} = \begin{pmatrix} 0.0153 & -0.6635 \\ -0.5088 & -0.0642 \\ -0.4650 & 0.1059 \end{pmatrix} \text{ und } F_{mis}^{(2)} = \begin{pmatrix} -0.3733 & -0.4947 \\ -0.5026 & -0.0787 \\ -0.3644 & 0.5418 \end{pmatrix}$$

ergibt sich auszugsweise: $\hat{\tilde{A}}_{mis} = \tilde{A}_{obs} \cdot F_{obs}^{(2)} \cdot F_{mis}^{(2)^T} \cdot \left(I - F_{mis}^{(2)} \cdot F_{mis}^{(2)^T} \right) = \begin{pmatrix} -0.9201 & -0.2147 & 0.8885 \\ -1.4252 & -1.0653 & 0.0979 \\ \vdots & \vdots & \vdots \\ 0.2759 & 0.9858 & 1.3411 \end{pmatrix}$.

Nach entsprechender Umkehrung der durchgeführten Standardisierung der Daten ergeben sich die in der nachfolgenden Tabelle dargestellten Imputationswerte für die fehlenden Daten. Dabei sind zusätzlich die Ergebnisse bei ausschließlicher Verwendung der ersten Hauptkomponente, bei Verwendung der ersten drei Hauptkomponenten sowie im Fall der Iteration (It.) des Verfahrens bei Verwendung der ersten beiden Hauptkomponenten angegeben.

Fehlende Ausprägung	a_{24}	$a_{12,4}$	$a_{14,4}$	$a_{12,7}$	$a_{14,7}$	$a_{15,7}$	a_{49}	$a_{11,9}$
Imputationswert mit $F^{(1)}$	67.69	81.26	70.88	57.96	27.85	75.67	18.24	44.29
Imputationswert mit $F^{(2)}$	57.80	84.85	65.90	59.57	25.63	72.49	6.64	55.13
Imputationswert mit $F^{(2)}$ (It.)	57.32	83.84	65.02	62.19	28.59	78.07	4.36	49.92
Imputationswert mit $F^{(3)}$	60.97	67.76	60.68	57.88	25.11	73.05	0.71	78.10

Im Hinblick auf die dargestellten Ergebnisse im Fall der iterativen Anwendung des Verfahrens ist noch anzumerken, daß es sich hierbei um die nach sieben Iterationen resultierenden Imputationswerte handelt. Weitere Iterationen führen bei der in diesem Beispiel verwendeten Genauigkeit von zwei bzw. vier Dezimalstellen zu keiner Veränderung dieser Imputationswerte.

Wie *Gleason und Staelin (1975, S. 234-236)* zeigen, führt die von ihnen vorgeschlagene Methode zu denselben Ergebnissen wie die entsprechende Anwendung einer multiplen Regression, falls die Information der m Ausgangsmerkmale vollständig auf die herangezogenen Hauptkomponenten übertragen werden kann und die Matrix $(I - F_{mis}^{(t)} \cdot F_{mis}^{(t)^T})$ regulär ist. Ist darüber hinaus die Anzahl der verwendeten Ladungsvektoren oder die Anzahl der Merkmale mit fehlenden Daten kleiner oder gleich der Anzahl der Merkmale ohne fehlende Daten, d.h. $t \leq q$ oder $(m-q) \leq q$, dann ist sogar eine exakte Reproduktion

der Matrix $\tilde{A}_{mis}$ möglich. Mit dieser zusätzlich geforderten Bedingung wird gewährleistet, daß der Rang der Matrix $\hat{\tilde{A}}_{mis}$ nicht kleiner als der Rang der Matrix $\tilde{A}_{mis}$ ist (vgl. *Gleason, Staelin, 1975, S. 235*).

4.2.3.5 Imputation auf Basis von Distanzeigenschaften

Bei Vorliegen einer unvollständigen, unmittelbar erhobenen Distanzmatrix können die Beziehungen zwischen einzelnen paarweisen Distanzen, die sich aufgrund von geforderten Distanzeigenschaften ergeben, zur Bestimmung von Imputationswerten herangezogen werden. Als Distanzeigenschaften werden die **ultrametrischen Ungleichung**

$$h,i,j \in N \;\Rightarrow\; d_{ij} \leq \max\left\{d_{ih}, d_{jh}\right\} \tag{4.39}$$

sowie die in (4.10) angegebene **Dreiecksungleichung** betrachtet. Die Dreiecksungleichung ist weniger restriktiv als die ultrametrische Ungleichung, da die ultrametrische Ungleichung die Dreiecksungleichung impliziert. Während im Fall einer Berechnung der paarweisen Distanzen aus einer Datenmatrix die meisten der merkmalsweisen Distanzindizes die Dreiecksungleichung oder sogar die ultrametrische Ungleichung erfüllen (vgl. z.B. *Opitz, 1980, S. 34, 36-37*), sind bei einer unmittelbar erhobenen Distanzmatrix diese Bedingungen im allgemeinen verletzt. Die Gültigkeit der Dreiecksungleichung oder der ultrametrischen Ungleichung ist zwar für die Datenanalyse nicht von entscheidender Bedeutung, jedoch stellt ein Heranziehen dieser Distanzeigenschaften eine einfache Möglichkeit zur Bestimmung von Imputationswerten dar. Die Idee geht dabei auf *de Soete (1984a, S. 238-239, 1984b, S. 389)* zurück, der aus einer unvollständigen Distanzmatrix eine vollständige Distanzmatrix schätzt, die der unvollständigen Distanzmatrix nach dem Kleinst-Quadrate-Prinzip am besten entspricht und deren paarweise Distanzen der ultrametrischen Ungleichung genügen. Der von *de Soete* beschriebene, iterative Algorithmus geht zwar über die hier beabsichtigte Bestimmung der Imputationswerte auf Basis einer Distanzeigenschaft hinaus, jedoch stellen die dabei vorgeschlagenen Anfangsschätzungen für die fehlenden Distanzen

$$d_{ij}^{\max} = \min_{h \in N,\, h \neq i,j} \left\{ \max_{i,j} \left\{ d_{ih}, d_{jh} \right\} \right\} \quad \forall\, i,j : w_{ij} = 0 \tag{4.40}$$

$$\text{mit} \begin{cases} d_{ih} = 0 & \text{falls } w_{ih} = 0 \wedge w_{jh} = 1 \\ d_{jh} = 0 & \text{falls } w_{ih} = 1 \wedge w_{jh} = 0 \\ d_{ih} = d_{jh} = \infty & \text{falls } w_{ih} = w_{jh} = 0 \end{cases}$$

Höchstwerte im Hinblick auf die Einhaltung der ultrametrischen Ungleichung dar. Verwendet man lediglich die Objekte $h \in N$, $h \neq i, j$, deren paarweise Distanzen bezüglich

der beiden Objekte i und j vorhanden sind, dann ergeben sich die Höchstwerte für die fehlenden Distanzen nach der Formel

$$d_{ij}^{\max} = \min_{\substack{h \in N,\, h \neq i,j, \\ h:\, w_{ih}=w_{jh}=1}} \left\{ \max_{i,j} \left\{ d_{ih}, d_{jh} \right\} \right\} \quad \forall\, i,j: w_{ij} = 0\,. \tag{4.41}$$

Die nach (4.41) berechneten Werte sind grundsätzlich größer oder gleich den nach (4.40) bestimmten Werten. Da die ultrametrische Ungleichung auch einzuhalten ist, wenn die fehlende Distanz auf der rechten Seite der Ungleichung steht, können analog Mindestwerte ermittelt werden. Diese ergeben sich, je nach verwendeter Datenbasis, gemäß

$$d_{ij}^{\min} = \max_{h \in N,\, h \neq i,j} \left\{ \max_{i,j} \left\{ d_{ih}, d_{jh} \right\} \right\} \quad \forall\, i,j: w_{ij} = 0 \tag{4.42}$$

$$\text{mit} \begin{cases} d_{ih} = 0 & \text{falls } w_{ih} = 0 \wedge w_{jh} = 1 \\ d_{jh} = 0 & \text{falls } w_{ih} = 1 \wedge w_{jh} = 0\,, \\ d_{ih} = d_{jh} = \infty & \text{falls } w_{ih} = w_{jh} = 0 \end{cases}$$

$$d_{ij}^{\min} = \max_{\substack{h \in N,\, h \neq i,j, \\ h:\, w_{ih}=w_{jh}=1}} \left\{ \max_{i,j} \left\{ d_{ih}, d_{jh} \right\} \right\} \quad \forall\, i,j: w_{ij} = 0 \tag{4.43}$$

Die nach (4.42) bzw. (4.43) berechneten Mindestwerte sind immer größer oder gleich den entsprechenden Höchstwerten. Sind also Höchst- und Mindestwert für eine fehlende Distanz d_{ij} identisch, d.h. $d_{ij}^{\max} = d_{ij}^{\min}$, dann sind im Fall einer Imputation mit diesem Wert alle nach (4.39) für die fehlende Distanz d_{ij} resultierenden Ungleichungen erfüllt. Ist der Höchstwert kleiner als der Mindestwert, dann existiert für die fehlende Distanz kein Imputationswert, der allen entsprechenden Ungleichungen genügt. In diesem Fall kann als Imputationswert jedoch das arithmetische Mittel aus Höchst- und Mindestwert herangezogen werden, da dadurch zumindest eine „gleichmäßige Abweichung" von der ultrametrischen Ungleichung gewährleistet wird. Berücksichtigt man zusätzlich die im Fall zufällig fehlender Distanzen plausible Forderung, daß der Mittelwert der ersetzten Distanzen gleich dem Mittelwert der vorhandenen Distanzen sein soll, dann bietet sich die Verwendung eines entsprechend gewichteten Mittels an. Mit

$$d_{ij} = \alpha \cdot d_{ij}^{\max} + (1-\alpha) \cdot d_{ij}^{\min} \quad \forall\, i,j: w_{ij} = 0, \tag{4.44}$$

$$\alpha = \frac{\dfrac{1}{w^{obs}} \cdot \sum\limits_{i,j:\, w_{ij}=1,\, i<j} d_{ij} - \dfrac{1}{w^{mis}} \cdot \sum\limits_{i,j:\, w_{ij}=0,\, i<j} d_{ij}^{\min}}{\dfrac{1}{w^{mis}} \cdot \sum\limits_{i,j:\, w_{ij}=0,\, i<j} d_{ij}^{\max} - \dfrac{1}{w^{mis}} \cdot \sum\limits_{i,j:\, w_{ij}=0,\, i<j} d_{ij}^{\min}}$$

ergibt sich gemäß diesem Ansatz der Imputationswert für eine fehlende Distanz.

Beispiel:

Für die fehlenden paarweisen Distanzen der Distanzmatrix des Anhangs B (Fall 1) erhält man, abhängig von der verwendeten Datengrundlage, die in der nachfolgenden Tabelle angegebenen Höchst- und Mindestwerte sowie die daraus resultierenden Imputationswerte auf Basis der ultrametrischen Ungleichung. In der rechten Spalte der Tabelle sind zusätzlich die arithmetischen Mittel dieser Werte dargestellt.

Fehlende Distanz	d_{14}	$d_{1,10}$	d_{27}	$d_{2,10}$	d_{35}	d_{37}	d_{46}	d_{59}	d_{69}	
Höchstwert nach (4.40)	3.95	2.57	1.92	1.92	2.06	2.06	2.08	2.35	2.08	2.33
Mindestwert nach (4.42)	7.88	7.88	9.43	9.43	9.43	9.43	5.65	7.18	7.18	8.17
Imputationswert nach (4.44)	5.34	4.45	4.58	4.58	4.67	4.67	3.34	4.06	3.89	4.40
Höchstwert nach (4.41)	4.25	3.95	4.25	4.45	3.58	3.55	3.02	2.35	3.95	3.71
Mindestwert nach (4.43)	7.88	7.88	7.18	9.43	9.43	7.88	5.65	7.18	7.18	7.74
Imputationswert nach (4.44)	4.87	4.62	4.75	5.30	4.58	4.29	3.47	3.18	4.50	4.40

Der Mittelwert der vorhandenen paarweisen Distanzen beträgt 4.40. Damit ergibt sich für eine Gewichtung der Höchst- und Mindestwerte im Rahmen der Bestimmung der Imputationswerte nach (4.44) der Wert $\alpha = (4.40 - 8.17)/(2.33 - 8.17) = 0.646$ bzw. $\alpha = (4.40 - 7.74)/(3.71 - 7.74) = 0.829$. Da alle nach (4.42) bzw. (4.43) bestimmten Mindestwerte über den jeweiligen Höchstwerten gemäß (4.40) bzw. (4.41) liegen, kann kein einziger Imputationswert für eine fehlende Distanz so gewählt werden, daß alle mit diesem Imputationswert nach (4.39) resultierenden Ungleichungen erfüllt sind.

Analog zur ultrametrischen Ungleichung muß die Einhaltung der Dreiecksungleichung für die beiden Fälle, daß die fehlende Distanz entweder auf der linken oder aber auf der rechten Seite der Ungleichung steht, gewährleistet sein. Entsprechend ergeben sich die Höchst- und Mindestwerte für die fehlenden paarweisen Distanzen gemäß

$$d_{ij}^{\max} = \min_{\substack{h \in N,\, h \neq i,j,\\ h:\, w_{ih} = w_{jh} = 1}} \left\{ d_{ih} + d_{jh} \right\} \quad \forall\, i,j: w_{ij} = 0\,, \tag{4.45}$$

$$d_{ij}^{\min} = \max_{\substack{h \in N,\, h \neq i,j,\\ h:\, w_{ih} = w_{jh} = 1}} \left\{ \left| d_{ih} - d_{jh} \right| \right\} \quad \forall\, i,j: w_{ij} = 0\,. \tag{4.46}$$

Aufgrund der additiven Komponente in der Dreiecksungleichung können lediglich die Objekte $h \in N$, $h \neq i, j$ herangezogen werden, deren paarweise Distanzen bezüglich der beiden Objekte i und j vorhanden sind. Im Gegensatz zu den auf Basis der ultrametrischen Ungleichung bestimmten Höchst- und Mindestwerten können die gemäß (4.46) ermittelten Mindestwerte kleiner als die entsprechenden Höchstwerte nach (4.45) sein. In diesem Fall sind bei Verwendung eines Imputationswerts aus dem Intervall [Mindestwert, Höchstwert] alle nach (4.10) für diesen Wert resultierenden Ungleichungen erfüllt. Die in (4.44) dargestellte Mittelwertbildung zur Bestimmung der Imputationswerte

für die fehlenden paarweisen Distanzen ist damit natürlich auch bei einer Imputation auf Basis der Dreiecksungleichung geeignet.

Beispiel:

Für die fehlenden paarweisen Distanzen der Distanzmatrix des Anhangs B (Fall 1) ergeben sich auf Basis der Dreiecksungleichung die folgenden Höchst-, Mindest- und daraus resultierenden Imputationswerte, wobei in der rechten Spalte der Tabelle wiederum die jeweiligen arithmetischen Mittel angegeben sind:

Fehlende Distanz	d_{14}	$d_{1,10}$	d_{27}	$d_{2,10}$	d_{35}	d_{37}	d_{46}	d_{59}	d_{69}	
Höchstwert nach (4.45)	6.90	6.17	6.81	6.54	7.13	5.60	5.67	4.43	9.08	6.48
Mindestwert nach (4.46)	3.15	2.87	3.27	3.08	4.10	4.43	2.61	5.12	4.16	3.64
Imputationswert nach (4.44)	4.16	3.75	4.22	4.01	4.91	4.74	3.43	4.94	5.48	4.40

Bei einem Mittelwert der vorhandenen paarweisen Distanzen von 4.40 erhält man im Hinblick auf die Gewichtung der Höchst- und Mindestwerte in (4.44) den Wert $\alpha = (4.40 - 3.64)/(6.48 - 3.64) = 0.268$. Mit einer Ausnahme sind alle Mindestwerte kleiner als die jeweiligen Höchstwerte. Damit sind die mit dem entsprechenden Imputationswert nach (4.10) resultierenden Ungleichungen erfüllt.

Die Bestimmung von Imputationswerten auf Basis der Dreiecksungleichung besitzt im Vergleich zu einer Imputation unter Verwendung der ultrametrischen Ungleichung zwar den Nachteil, daß nicht alle vorhandenen paarweisen Distanzen herangezogen werden können, dafür genügen die resultierenden Imputationswerte im allgemeinen jedoch der zugrundegelegten Bedingung und sind daher plausibler.

4.2.4 Imputation bei systematischen Ausfallmechanismen

Bei Vorliegen eines systematischen Ausfallmechanismus ist eine adäquate Behandlung der fehlenden Daten nur dann möglich, wenn der Ausfallmechanismus bekannt ist. In diesem Fall sind dann Verfahren zur Behandlung der fehlenden Daten heranzuziehen, die ein Modell des Ausfallmechanismus verwenden. Entsprechende Imputationsverfahren auf Basis eines sogenannten **Ausfallmodells** können sowohl im Fall einer unvollständigen Datenmatrix, wie auch im Fall einer unvollständigen, unmittelbar erhobenen Distanzmatrix angewandt werden. Grundsätzlich existieren jedoch nur wenige Möglichkeiten einer derartigen Imputation. Als Imputationswerte kommen entweder ein Lageparameter oder zufällig ausgewählte Werte unter Berücksichtigung des Ausfallmodells in Betracht. Der jeweils heranzuziehende Lageparameter hängt dabei vom zugrundeliegenden Skalenniveau ab. Im Fall eines kardinalen Merkmals oder bei Vorliegen einer Distanzmatrix stellt der Erwartungswert, im Fall eines ordinalen Merkmals der Median und im Fall eines nominalen Merkmals der Modus einen geeigneten Lageparameter dar.

Im folgenden werden diese Ansätze lediglich im Zusammenhang mit einer unvollständige Datenmatrix betrachtet. Eine Imputation im Fall einer unvollständigen, unmittelbar erhobenen Distanzmatrix kann jedoch analog durchgeführt werden.

In Anlehnung an *Greenless et al. (1982, S. 253)* ergibt sich der **Erwartungswert** für eine fehlende Ausprägung a_{ik} ($i \in N$, $k \in M$) eines kardinal skalierten Merkmals k gemäß

$$E\left(a_{ik} \mid A, v_{ik}=0\right)=\begin{cases}\dfrac{\int_{-\infty}^{+\infty} a_{ik} \cdot P\left(v_{ik}=0 \mid A\right) da_{ik}}{\int_{-\infty}^{+\infty} P\left(v_{ik}=0 \mid A\right) da_{ik}} & \text{falls } a_{ik} \text{ stetig} \\ \dfrac{\sum_{a_{ik}} a_{ik} \cdot P\left(v_{ik}=0 \mid A\right)}{\sum_{a_{ik}} P\left(v_{ik}=0 \mid A\right)} & \text{falls } a_{ik} \text{ diskret}\end{cases} . \tag{4.47}$$

Dabei stellt die Ausprägung a_{ik} einerseits eine Zufallsvariable (linke Seite der Gleichung) und andererseits deren mögliche Realisierungen, d.h. alle denkbaren Merkmalsausprägungen des Merkmals k (rechte Seite der Gleichung) dar. $P\left(v_{ik}=0 \mid A\right)$ bezeichnet die bedingte Wahrscheinlichkeit, daß die Ausprägung a_{ik} bei gegebener Datenmatrix A fehlt. Der nach (4.47) bestimmte Erwartungswert kann dann als Imputationswert für die fehlende Ausprägung a_{ik} herangezogen werden. Besitzt das Merkmal k nominales oder ordinales Skalenniveau, dann kann der Modus bzw. der Median als Imputationswert herangezogen werden.

Beispiel:

Für die Datenmatrix des Anhangs A wird für den dort beschriebenen Fall 2, der die Abhängigkeit des Fehlens der Daten beim Merkmal Multivariate Verfahren (6) von den tatsächlichen Realisierung dieser Werte berücksichtigt, das folgende Modell für den zugrundeliegenden Ausfallmechanismus unterstellt:

$$P\left(v_{i6}=0 \mid A\right)=\begin{cases}0.2 & \text{für } 0 \le a_{i6} \le 50 \\ 0.9 & \text{für } 50 < a_{i6} \le 100 \\ 0 & \text{sonst}\end{cases} .$$

Das Merkmal Multivariate Verfahren ist stetig und damit ergibt sich unter Verwendung des entsprechenden Astes in (4.47) der folgende Erwartungswert:

$$E\left(a_{i6} \mid A, v_{i6}=0\right)=\frac{\int_0^{50} a_{i6} \cdot 0.2\, da_{i6}+\int_{50}^{100} a_{i6} \cdot 0.9\, da_{i6}}{\int_0^{50} 0.2\, da_{i6}+\int_{50}^{100} 0.9\, da_{i6}}=\frac{0.1 \cdot \left(a_{i6}\right)^2\Big|_0^{50}+0.45 \cdot \left(a_{i6}\right)^2\Big|_{50}^{100}}{0.2 \cdot a_{i6}\Big|_0^{50}+0.9 \cdot a_{i6}\Big|_{50}^{100}}=\frac{250+3375}{10+45}=65.91$$

Dieser Wert kann schließlich als Imputationswert für die fehlenden Ausprägungen des Merkmals Multivariate Verfahren herangezogen werden.

Für den in Anhang A beschriebenen Fall 4, der die Abhängigkeit der fehlenden Ausprägungen beim Merkmal Preisniveau (1) von den Werten der Ausprägungen des Merkmals Deskriptive Statistik (4) betrachtet, stellt aufgrund des ordinalen Skalenniveaus des Merkmals Preisniveau der Median einen geeigneten Lageparameter dar. Die Bildung des Erwartungswerts gemäß (4.47) ist aufgrund des Skalenniveaus eigentlich nicht möglich, soll aber neben der Bestimmung des Medians zur beispielhaften Darstellung der Methodik durchgeführt werden. Die Ausprägungen des Merkmals Preisniveau werden dazu folgendermaßen codiert: niedrig = 1, mittel = 2, gehoben = 3, hoch = 4. Für das Fehlen einer Merkmalsausprägung wird die Wahrscheinlichkeit

$$P(v_{i1}=0 \mid A)=\begin{cases}0.1 & \text{für } 0 \le a_{i4} \le 70\\ 0.7 & \text{für } 70 < a_{i4} \le 100\\ 0 & \text{sonst}\end{cases}$$

unterstellt. Des weiteren erhält man aufgrund der ursprünglich vorliegenden, vollständigen Daten die folgende, als bekannt vorausgesetzte Häufigkeitsverteilung:

		Preisniveau			
		1 (niedrig)	**2 (mittel)**	**3 (gehoben)**	**4 (hoch)**
Deskriptive Statistik	**[0;70]**	2	2	0	1
	⟨70;100]	0	4	2	4

Das zugrundeliegende Modell für den Ausfallmechanismus sowie der damit resultierende Erwartungswert für die Ausprägungen des Merkmals Preisniveau ergeben sich schließlich gemäß

$$P(v_{i1}=0 \mid A)=\begin{cases}\frac{2}{2}\cdot 0.7+\frac{0}{2}\cdot 0.1=0.7 & \text{für } a_{i1}=1\\ \frac{2}{6}\cdot 0.7+\frac{4}{6}\cdot 0.1=0.3 & \text{für } a_{i1}=2\\ \frac{0}{2}\cdot 0.7+\frac{2}{2}\cdot 0.1=0.1 & \text{für } a_{i1}=3\\ \frac{1}{5}\cdot 0.7+\frac{4}{5}\cdot 0.1=0.22 & \text{für } a_{i1}=4\\ 0 & \text{sonst}\end{cases},$$

$$E(a_{i1} \mid A, v_{i1}=0)=\frac{1\cdot 0.7+2\cdot 0.3+3\cdot 0.1+4\cdot 0.22}{0.7+0.3+0.1+0.22}=\frac{2.48}{1.32}=1.88\,.$$

Eine Verwendung dieses Erwartungswerts als Imputationswert sollte jedoch aufgrund des vorliegenden ordinalen Skalenniveaus und des somit von der gewählten Codierung abhängigen Ergebnisses nicht erfolgen. Der im Hinblick auf das Skalenniveau geeignete Lageparameter ist der Median, der sich wegen $\frac{0.7}{0.7+0.3+0.1+.0.22}=0.53$ mit einem Wert von 1 bzw. niedrig ergibt.

Eine Imputation mittels **Zufallsauswahl** auf Basis eines Modells des Ausfallmechanismus kann unter Berücksichtigung der in Abschnitt 4.2.1.3 gemachten Ausführungen erfolgen. Dabei stellt sowohl eine im Hinblick auf das Ausfallmodell aus einem Zufallszahlengenerator gezogene Zahl, wie auch eine im Hinblick auf das Ausfallmodell aus den vorhandenen Daten zufällig ausgewählte Merkmalsausprägung einen möglichen Imputationswert dar.

Beispiel:

Für die Datenmatrix des Anhangs A wird im folgenden der Fall 2 betrachtet. Ausgehend von dem im vorherigen Beispiel bereits dargestellten Ausfallmodell für die fünf zu bestimmenden Imputationswerte ergibt sich ein Verhältnis von 0.2 : 0.9 im Hinblick auf die aus den Intervallen [0;50] und ⟨50;100] auszuwählenden Werte, d.h. 18 Prozent der fünf benötigte Werte, also gerundet ein Wert ist dem Intervall [0;50] und 82 Prozent der fünf benötigten Werte, also gerundet vier Werte sind dem Intervall ⟨50;100] zu entnehmen. Innerhalb der beiden Intervalle können die jeweils benötigten Werte dann entweder aus einem Zufallszahlengenerator gezogen oder aber aus den vorhandenen Ausprägungen zufällig ausgewählt werden. Die auf diese Art bestimmten Imputationswerte sind schließlich zufällig auf die fehlenden Ausprägungen zu verteilen.

4.2.5 Vergleich und Kombination der Verfahren

Neben den grundsätzlichen Voraussetzungen der einzelnen Verfahren, dazu zählen das Vorliegen einer Daten- oder Distanzmatrix, das Skalenniveau der Merkmale und der zugrundeliegende Ausfallmechanismus, unterscheiden sich die vorgestellten Imputationstechniken vor allem hinsichtlich des Aufwands, der mit der Anwendung der einzelnen Verfahren verbunden ist, sowie der zur Schätzung der fehlenden Daten verwendeten Informationen. In der Tabelle 4.2 sind die Voraussetzungen sowie die wichtigsten Eigenschaften der vorgestellten Imputationsverfahren noch einmal zusammenfassend dargestellt.

Eine Kombination mehrerer Imputationstechniken wird vor allem im Fall gemischter Datenmatrizen zur Anwendung kommen. So können beispielsweise die fehlenden Daten kardinaler Merkmale mittels der vorgestellten Regressionsmethoden geschätzt werden, während eine Ersetzung der fehlenden Ausprägungen nominaler und ordinaler Merkmale mittels der Diskriminanzanalyse oder des entsprechenden Lageparameters erfolgt.

Eine weitere Kombinationsmöglichkeit von Imputationsverfahren findet sich bei *Little und Rubin (1987, S. 61)*. Dabei werden zunächst unter Verwendung einer Regressionsmethode Schätzwerte für die fehlenden Daten bestimmt. Anschließend wird zu jedem Schätzwert ein empirisches Residuum addiert. Dieses wird mittels eines Hot-Deck-Verfahrens aus den Residuen, die sich auf Basis des verwendeten Regressionsmodells ergeben, ausgewählt. Die auf diese Art korrigierten Schätzwerte werden dann als Imputationswerte für die fehlenden Daten herangezogen.

Alle vorgestellten Imputationstechniken, ob einzeln oder in Kombination betrachtet, führen letztendlich zu einer vollständigen Daten- bzw. Distanzmatrix, die mit den bekannten datenanalytischen Methoden ausgewertet werden kann.

Imputation mittels	Voraussetzungen	Eigenschaften
Lageparameter	Daten- oder Distanzmatrix, Daten sind MCAR	Einfache Anwendbarkeit, Informationen der anderen Merkmale werden nicht berücksichtigt
Verhältnisschätzer	Datenmatrix (kardinales Merkmal mit MD, kardinales Hilfsmerkmal), Daten sind MCAR	Einfache Anwendbarkeit, Informationen des Hilfsmerkmals werden berücksichtigt
Zufallsauswahl	Daten- oder Distanzmatrix, Daten sind MCAR	Verteilung der Merkmalsausprägungen bzw. Distanzen muß bekannt sein oder festgelegt werden
Expertenratings	Daten- oder Distanzmatrix, Daten sind MCAR	Hoher Aufwand, unter Umständen unzutreffende Zusammenhangs- oder Kausalannahmen des Experten
Cold-Deck-Verfahren	Datenmatrix, Daten sind MCARC	Berücksichtigung der Ähnlichkeitsbeziehungen der Objekte, externe Informationen werden benötigt
Hot-Deck-Verfahren	Datenmatrix, Daten sind MCARC	Berücksichtigung der Ähnlichkeitsbeziehungen der Objekte, Verdopplung vorhandener Daten
Regressionsanalyse	Datenmatrix (kardinale und dichotome Merkmale, kardinale Merkmale mit MD), Daten sind MCAR oder MAR	Informationen anderer Merkmale werden berücksichtigt, Beschränkungen hinsichtlich des Skalenniveaus der Merkmale
Varianzanalyse	Datenmatrix (kardinale und nominale Merkmale, kardinale Merkmale mit MD), Daten sind MCAR oder MAR	Informationen anderer Merkmale werden berücksichtigt, Beschränkungen hinsichtlich des Skalenniveaus der Merkmale
Diskriminanzanalyse	Datenmatrix (kardinale und nominale Merkmale, nominale Merkmale mit MD), Daten sind MCAR oder MAR	Informationen anderer Merkmale werden berücksichtigt, Beschränkungen hinsichtlich des Skalenniveaus der Merkmale
Hauptkomponentenmethode	Datenmatrix (quantitativ), Daten sind MCAR oder MAR	Informationen anderer Merkmale werden berücksichtigt, Beschränkungen hinsichtlich des Skalenniveaus der Merkmale
Distanzeigenschaften	Distanzmatrix, Daten sind MCAR	Beziehungen zwischen einzelnen paarweisen Distanzen werden unterstellt bzw. berücksichtigt
Ausfallmodellen	Daten- oder Distanzmatrix, Modell des Ausfallmechanismus	Modell des Ausfallmechanismus muß bekannt sein, Daten können systematisch fehlen

Tabelle 4.2: Vergleich der Imputationsverfahren

4.3 Parameterschätzverfahren

In der Literatur (vgl. z.B. *Schwab, 1991, S. 4*) werden unter dem Begriff Parameterschätzverfahren alle Methoden zusammengefaßt, die auf Basis einer unvollständigen Datenmatrix bestimmte Parameter, wie beispielsweise Mittelwerte, Varianzen, Kovarianzen, Regressionskoeffizienten oder Faktorladungen, schätzen. Mit dieser Begriffsdefinition ist jedoch keine deutliche Abgrenzung der Parameterschätzverfahren von den bereits vorgestellten Eliminierungs- und Imputationsverfahren sowie den im folgenden Abschnitt 4.4 noch zu behandelnden, multivariaten Analyseverfahren möglich.

Die folgende Betrachtung beschränkt sich daher lediglich auf Methoden, bei denen auf Basis einer unvollständigen Datenmatrix eine unmittelbare Schätzung von Verteilungsparametern erfolgt. Dabei wird in **Abschnitt 4.3.1** zunächst ein Überblick der aus der Literatur bekannten Schätzmethoden gegeben. Da die Verfahren entweder sehr spezielle Anforderungen an das vorliegende Datenmaterial stellen oder die zu schätzenden Parameter für eine Multivariate Datenanalyse von untergeordneter Bedeutung sind, wird auf eine ausführliche Darstellung aller angesprochenen Ansätze verzichtet. Lediglich der sogenannte EM-Algorithmus soll aufgrund seiner bedeutenden Stellung unter den Parameterschätzverfahren sowie seiner durchaus relevanten praktischen Anwendung in **Abschnitt 4.3.2** eingehender vorgestellt werden. In **Abschnitt 4.3.3** werden die angesprochenen Parameterschätzverfahren schließlich noch einmal miteinander verglichen und die in Betracht zu ziehenden Kombinationsmöglichkeiten einzelner Verfahren kurz diskutiert.

In diesem Abschnitt werden somit keine Schätzmethoden betrachtet, bei denen die Parameter aus den vollständig vorliegenden Objekten oder aus der vervollständigten Datenmatrix bestimmt werden. Darüber hinaus wird die Schätzung von Parametern, die das Ergebnis eines multivariaten Analyseverfahrens darstellen, in dieser Arbeit ebenfalls nicht unter dem Begriff der Parameterschätzverfahren eingeordnet. Dies ist damit zu begründen, daß eine derartige Schätzung entweder auf Basis von bereits geschätzten Verteilungsparametern erfolgt, also beispielsweise die Schätzung von Regressionskoeffizienten oder Faktorladungen auf einer bereits geschätzten Korrelationsmatrix basiert, oder aber auf ein multivariates Analyseverfahren, das die fehlenden Daten unmittelbar berücksichtigt, zurückzuführen ist.

Aufgrund der getroffenen Abgrenzung werden einige Methoden, die in der Literatur zu den Parameterschätzverfahren zählen, in diesem Abschnitt nicht behandelt. Dazu zählen beispielsweise die **Methode von *Glasser*** *(1964)*, bei der die Regressionskoeffizienten unter Verwendung der Mittelwerte und Varianzen bzw. Kovarianzen auf Basis der verfügbaren bzw. paarweise verfügbaren Objekte bestimmt werden, und die **Methode**

von *Yates* *(1933)*, bei der die Regressionskoeffizienten aus einer zuvor über Regressionsmethoden vervollständigten Datenmatrix geschätzt werden. ***Barlett's* ANCOVA Methode** *(1937)* wird hier ebenfalls nicht behandelt. Bei diesem Verfahren werden alle fehlenden Daten zunächst durch den Wert Null ersetzt. In das Regressionsmodell werden dann zusätzlich die MD-Indikatoren als Kovariablen aufgenommen, um somit aufgrund der zusätzlichen Modellparameter eine unverzerrte Schätzung der Störvarianz zu erhalten (vgl. *Toutenburg, 1992, S. 206*).

4.3.1 Überblick der Schätzmethoden

Die in der Literatur beschriebenen Parameterschätzverfahren lassen sich nach der zugrundeliegenden Theorie sowie nach dem vorausgesetzten Skalenniveau der Merkmale klassifizieren. Im Hinblick auf die zugrundeliegende Theorie ist dabei eine Unterscheidung in Verfahren auf Basis der Maximum-Likelihood-Theorie, Verfahren auf Basis der Bayes-Theorie sowie Verfahren ohne Verteilungsannahmen möglich. Im Hinblick auf das vorausgesetzte Skalenniveau können Verfahren für stetige Daten und Verfahren für kategorische Daten unterschieden werden. Während mit den Verfahren für stetige Daten die Mittelwerte, Varianzen und Kovarianzen geschätzt werden, erfolgt bei den Verfahren für kategorische Daten im Prinzip eine Parameterschätzung in Kontingenztabellen, d.h. auf Basis der für nominale Merkmale mit fehlenden Daten aufgestellten Kontingenztabellen werden in erster Linie die erwarteten Zellhäufigkeiten geschätzt. Da die Parameterschätzverfahren für kategoriale Daten lediglich zu einem sehr speziellen, für multivariate statistische Verfahren nicht weiter verwertbaren Analyseergebnis führen, werden im folgenden ausschließlich Verfahren für stetige Daten betrachtet. Dabei wird eine Einteilung nach der jeweils zugrundeliegenden Theorie vorgenommen. Eine Darstellung der Parameterschätzung in Kontingenztabellen kann beispielsweise *Blumenthal (1968)*, *Hocking und Oxspring (1971)*, *Fienberg (1972)*, *Koch et al. (1972)*, *Chen und Fienberg (1974, 1976)*, *Fuchs (1982)* sowie *Baker und Laird (1988)* entnommen werden.

Die Verfahren auf Basis der Maximum-Likelihood- bzw. der Bayes-Theorie setzen die Bedingung MAR für die Daten voraus. Die Verfahren ohne Verteilungsannahmen benötigen grundsätzlich die MCAR-Annahme und kommen lediglich bei Berücksichtigung sämtlicher Abhängigkeitsbeziehungen der fehlenden Werte von den vorhandenen Daten mit der MAR-Annahme aus.

4.3.1.1 Verfahren auf Basis der Maximum-Likelihood-Theorie

Die in diesem Abschnitt betrachteten Parameterschätzverfahren auf Basis der Maximum-Likelihood-Theorie gehen alle von einer quantitativen Datenmatrix aus. Je nach

Anzahl der vorliegenden Merkmale wird dabei eine bi- oder multivariate Normalverteilung unterstellt. Das Ziel dieser Verfahren besteht dann in der Schätzung der Verteilungsparameter, also der Mittelwerte, Varianzen und Kovarianzen der Merkmale.

Die ersten Ansätze einer Parameterschätzung auf Basis der Maximum-Likelihood-Theorie stellen die Arbeiten von ***Wilks** (1932)*, ***Matthai** (1951)*, ***Lord** (1955)* und ***Edgett** (1956)* dar. *Wilks* betrachtet die Maximum-Likelihood-Schätzer der Verteilungsparameter lediglich im Fall zweier Merkmale und *Matthai* erweitert diesen Ansatz auf drei Merkmale. Eine explizite Lösung der Maximum-Likelihood-Gleichungen zur Schätzung der Mittelwerte, Varianzen und Kovarianzen ist bei diesen Ansätzen jedoch nur dann möglich, wenn entweder die Mittelwerte oder aber die Varianzen und Kovarianzen bekannt sind. Deshalb gehen *Lord* und *Edgett* von speziellen Mustern fehlender Daten aus, für die eine explizite Lösung der Maximum-Likelihood-Gleichungen erfolgen kann. *Edgett* betrachtet dabei drei Merkmale, von denen aber lediglich eines fehlende Daten aufweist. *Lord* geht ebenfalls von insgesamt drei Merkmalen aus, wobei zwei Merkmale fehlende Daten in der Art besitzen, daß die Ausprägungen jeweils entweder für das eine oder für das andere Merkmal, aber nicht gleichzeitig für beide Merkmale fehlen oder vorhanden sind. Die Maximum-Likelihood-Schätzungen nach *Wilks*, *Matthai*, *Lord* und *Edgett* stellen damit alle sehr restriktive Anforderungen an das vorliegende Datenmaterial, so daß eine Anwendung dieser Ansätze stark eingeschränkt ist.

Eine Verallgemeinerung der Ansätze von *Lord* und *Edgett* wird durch die von *Anderson (1957)* vorgeschlagene **Faktorisierung der Likelihood** im Fall sogenannter **nested Muster** der fehlenden Daten ermöglicht.[11] Ein nested MD-Muster kann entweder **monoton** oder **nonmonoton** sein. Die Abbildung 4.4 zeigt die grundsätzliche Struktur einer unvollständigen Datenmatrix für diesen beiden Fälle eines nested MD-Musters (vgl. *Anderson, 1957, S. 202-203*, *Rubin, 1974, S. 468*). Die dunkel schraffierten Bereiche zeigen dabei vorhandene Daten, die hell schraffierten Bereiche teilweise vorhandene Daten und die nicht schraffierten Bereiche fehlende Daten an.

Das bei *Edgett* bzw. *Lord* vorausgesetzte Muster der fehlenden Daten stellt damit einen Spezialfall eines nested MD-Musters dar. Durch die Faktorisierung der Likelihood, d.h. die Maximum-Likelihood-Funktion aller Verteilungsparameter kann als Produkt mehrerer Funktionen einzelner Verteilungsparameter dargestellt werden, kann das Ausgangsschätzproblem in mehrere kleine Schätzprobleme zerlegt werden. Eine ausführliche Darstellung der Parameterschätzung auf Basis einer Faktorisierung der Likelihood ist bei *Little und Rubin (1987, S. 97-126)* zu finden. Die entsprechenden Maximum-Likelihood-Schätzungen für die Mittelwerte und die Kovarianzmatrix ergeben sich im Prinzip

[11] Der Begriff „nested" geht dabei auf *Hartley und Hocking (1971, S. 786)* zurück.

dadurch, daß die Datenmatrix zunächst auf Basis von Regressionsmodellen, die im Hinblick auf das Muster der fehlenden Daten geeignet sind, vervollständigt wird. Anschließend werden die Mittelwerte und die Kovarianzmatrix berechnet. Die Varianzen bzw. Kovarianzen der Merkmale mit fehlenden Daten sind schließlich noch durch Addition der entsprechenden, aus den Regressionsmodellen resultierenden Varianzen bzw. Kovarianzen der Residuen zu korrigieren (vgl. *Little, Rubin, 1987, S. 109, 116-117, 122-123*). Aufgrund der vorausgesetzten Muster der fehlenden Daten ist dieser Ansatz jedoch nur eingeschränkt anwendbar.

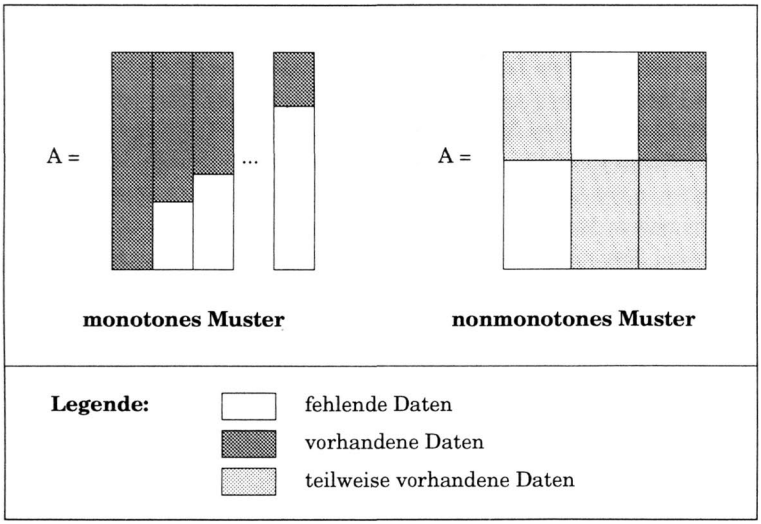

Abbildung 4.4: Struktur der Datenmatrix bei einem nested MD-Muster

Falls kein nested Muster der fehlenden Daten vorliegt, kann eine Maximierung der Likelihood-Funktion, die dann nicht faktorisiert werden kann, nur über iterative Verfahren erreicht werden (vgl. *Anderson et al. 1983, S. 458, Toutenburg, 1992, S. 214*). *Little und Rubin (1987, S. 127-129)* geben einen Überblick der in der Literatur bekannten Algorithmen zur iterativen Bestimmung der Maximum-Likelihood-Schätzungen für die Mittelwerte und die Kovarianzmatrix, wobei in diesem Zusammenhang in erster Linie der **EM-Algorithmus** sowie der sogenannte **Scoring-Algorithmus** zu nennen sind. Da der EM-Algorithmus sowohl im Hinblick auf das zugrundeliegende Konzept, wie auch im Hinblick auf die praktische Umsetzung den einfachsten Lösungsansatz zur Bestimmung von Maximum-Likelihood-Schätzungen für die Mittelwerte und die Kovarianzmatrix im Fall beliebiger Muster fehlender Daten darstellt, wird dieses Verfahren in Abschnitt 4.3.2 noch ausführlicher vorgestellt. Eine Anwendung des Scoring-Algorith-

mus kann beispielsweise *Hartley und Hocking (1971)*, *Hocking und Marx (1979)* sowie *Hocking (1983)* entnommen werden.

Die bislang angesprochenen Ansätze einer Maximum-Likelihood-Schätzung der Verteilungsparameter können auch im Hinblick auf systematische Ausfallmechanismen modifiziert werden. Dazu ist ein Modell des Ausfallmechanismus heranzuziehen und die mit diesem Modell abgebildeten Abhängigkeitsbeziehungen der fehlenden Daten gehen in die Maximum-Likelihood-Funktion ein. Ein Überblick dieser Ansätze kann beispielsweise *Little und Rubin (1987, S. 218-243)* entnommen werden.

4.3.1.2 Verfahren auf Basis der Bayes-Theorie

Die Schätzmethoden auf Basis der Bayes-Theorie gehen von einer quantitativen Datenmatrix sowie einer multivariaten Normalverteilung aus. Zusätzlich werden Vorinformationen über die zu schätzenden Parameter auf Grundlage des vorliegenden Stichprobenmechanismus herangezogen. Mit diesen a priori bekannten Informationen über die aus der Grundgesamtheit ausgewählten Objekte und den mit den vorhandenen Daten gegebenen Informationen kann dann eine geeignete Schätzfunktion zur Bestimmung der Verteilungsparameter gebildet werden. Eine ausführliche Darstellung von Parameterschätzverfahren auf Basis der Bayes-Theorie wird beispielsweise von *Press und Scott (1974)*, *Little (1982, S. 239-241)* sowie *Little und Rubin (1987, S. 247-255)* gegeben.

Auch die Parameterschätzverfahren auf Basis der Bayes-Theorie können im Hinblick auf systematische Ausfallmechanismen modifiziert werden, indem wiederum ein Modell des Ausfallmechanismus in das Ausgangsmodell zur Schätzung der Verteilungsparameter aufgenommen wird. Eine Darstellung dieses Ansatzes ist bei *Little und Rubin (1987, S. 259-264)* zu finden.

4.3.1.3 Verfahren ohne Verteilungsannahmen

In Abschnitt 4.2.3.1 wurde die Methode von *Buck* zur Bestimmung von Imputationswerten mittels der Regressionsanalyse bereits vorgestellt. Auf Basis der mit dieser Methode vervollständigten Datenmatrix schlägt ***Buck*** *(1960, S. 303-304)* nun die folgende Schätzung für die Kovarianzmatrix vor: Aus der vervollständigten Datenmatrix wird die Kovarianzmatrix zunächst berechnet. Anschließend erfolgt eine Korrektur der Kovarianzmatrix. Dabei werden zu allen Varianzen von Merkmalen mit fehlenden Daten sowie zu allen Kovarianzen von jeweils zwei Merkmalen mit fehlenden Daten die entsprechenden, aus den Regressionsmodellen resultierenden Varianzen bzw. Kovarianzen der Residuen addiert. Der Anteil der für das Merkmal bzw. für beide Merkmale jeweils gleichzeitig

fehlenden Daten ist bei dieser Korrektur zu berücksichtigen. Wie *Beale und Little (1975, S. 137)* anmerken, ist eine iterative Anwendung des Ansatzes von *Buck*, d.h. wiederholte Regressionsschätzung der fehlenden Daten auf Basis der im vorhergehenden Schritt bestimmten Kovarianzmatrix mit anschließender Neuschätzung der Kovarianzmatrix bis zum Erreichen von Konvergenz für die geschätzten Parameter, identisch zum EM-Algorithmus. Jedoch ist keine Verteilungsannahme notwendig und die Schätzwerte stellen damit keine Maximum-Likelihood-Schätzungen dar.

Weitere verteilungsfreie Parameterschätzverfahren stellen die sogenannten **Gewichtungsmethoden** dar, die zur Schätzung von Mittelwerten bei kardinalen Merkmalen herangezogen werden. Die Schätzwerte für die Merkmalsmittelwerte ergeben sich dabei durch eine geeignete Gewichtung der für jedes Merkmal vorhandenen Ausprägungen. Als objektspezifisches Gewicht wird die Wahrscheinlichkeit einer Auswahl des Objekts in der Ausgangsstichprobe oder zusätzlich sogar die Wahrscheinlichkeit für das Vorhandensein der entsprechenden Merkmalsausprägung herangezogen. Zur Vereinfachung werden diese Auswahl- und Ausfallwahrscheinlichkeiten meist nur für die sogenannten Gewichtungsklassen, d.h. Klassen von ähnlichen Objekten bestimmt. Aufgrund der eher untergeordneten Bedeutung einer ausschließlichen Schätzung von Mittelwerten in einer datenanalytischen Untersuchung sowie der im Hinblick auf den Stichprobenmechanismus benötigten Informationen wird auf eine ausführliche Darstellung der Gewichtungsmethoden verzichtet und statt dessen auf die Arbeiten von *Chapman (1976, S. 247-249)*, *Oh und Scheuren (1983)*, *Platek und Gray (1983, S. 262, 270)*, *Madow et al. (1983, S.82-83)* sowie *Little und Rubin (1987, S. 55-60)* verwiesen.

4.3.2 EM-Algorithmus

Der EM-Algorithmus (expectation maximization algorithm) stellt ein sehr universelles, iteratives Verfahren zur Maximum-Likelihood-Schätzung von Parametern im Fall unvollständiger Datenmatrizen dar, wobei die Anwendung dieses Algorithmus nicht ausschließlich auf das Problem fehlender Daten beschränkt ist.[12] Die erste Arbeit einer Anwendung des EM-Algorithmus zur Parameterschätzung bei fehlenden Daten geht auf *Orchard und Woodbury (1972)* zurück, die allerdings von einem sogenannten **missing information principle** sprechen. Der Begriff „EM-Algorithmus" wurde erst von *Dempster et al. (1977)* geprägt, die auf Basis der Arbeit von *Orchard und Woodbury* den Algorithmus mit seinen Eigenschaften und Anwendungsmöglichkeiten allgemeingültig darstellen. Im folgenden sollen die theoretischen Grundlagen sowie die Anwendung des EM-

[12] Einen kurzen Überblick einiger Anwendungsmöglichkeiten des EM-Algorithmus geben beispielsweise *Little und Rubin (1987, S. 129)*.

Algorithmus zur Schätzung der Verteilungsparameter einer unvollständigen quantitativen Datenmatrix, deren Ausprägungen einer multivariaten Normalverteilung genügen, ausführlich vorgestellt werden.

Ausgehend von einer Verteilung für die Datenmatrix A mit dem Parameter θ und der zugehörigen Dichtefunktion $f(A|\theta) = f(A^{obs}, A^{mis}|\theta)$ ist gemäß dem Maximum-Likelihood-Prinzip ein Schätzwert $\hat{\theta}$ für θ gesucht, so daß die Dichtefunktion der vorhandenen Daten $f(A^{obs}|\theta)$ maximiert wird. Eine unmittelbare Bestimmung des Schätzwerts nach diesem Ansatz ist im allgemeinen jedoch schwierig. Einfacher und, wie nachfolgend noch gezeigt wird, äquivalent zur Maximierung von $f(A^{obs}|\theta)$ ist die Ermittlung eines Werts für θ, der die Likelihood-Funktion $f(A|\theta)$ maximiert, sofern die Matrix A vervollständigt ist, d.h. die fehlenden Daten zuvor ersetzt werden. Aufgrund der unter der MAR-Annahme geltenden Beziehung

$$f(A|\theta) = f(A^{obs}|\theta) \cdot f(A^{mis}|A^{obs},\theta) \tag{4.48}$$

$$\Leftrightarrow \quad \ln f(A|\theta) = \ln f(A^{obs}|\theta) + \ln f(A^{mis}|A^{obs},\theta)$$

müssen die Imputationswerte für die fehlenden Daten auf Basis der bedingten Verteilung von A^{mis} bei gegebenem A^{obs} und θ bestimmt werden. Zieht man einen beliebigen Parameter $\theta^{(0)}$ für θ heran, dann kann der Erwartungswert für beide Seiten von (4.48) bestimmt werden, wobei sich im Fall der logarithmierten Likelihood-Funktionen die Erwartungswertbildung wie folgt ergibt:

$$E\left[\ln f(A|\theta) \mid A^{obs},\theta^{(0)}\right] = \ln f(A^{obs}|\theta) + E\left[\ln f(A^{mis}|A^{obs},\theta) \mid A^{obs},\theta^{(0)}\right]. \tag{4.49}$$

Gesucht ist nun ein Wert θ^* für den Parameter θ, der die linke Seite der Gleichung (4.49) maximiert. Dieser Wert wird gemäß dem Ansatz von *Orchard und Woodbury* durch Nullsetzen der Ableitung dieser Likelihood-Funktion nach θ bestimmt. Wie *Beale und Little (1975, S. 132)* zeigen, wird dadurch implizit eine Funktion g definiert, die den Wert $\theta^{(0)}$ in den Wert θ^* transformiert, d.h.

$$\theta^* = g(\theta^{(0)}). \tag{4.50}$$

Der Parameter θ ist dann als Fixpunkt der Transformation g zu schätzen, d.h. ein Wert für θ ist so zu bestimmen, daß die Gleichung

$$\theta = g(\theta) \tag{4.51}$$

erfüllt ist. Falls die Likelihood-Funktion differenzierbar ist, genügt jede Lösung von (4.51) der gleich Null gesetzten Ableitung der Likelihood-Funktion der linken Seite von

(4.49). Da der zweite Term auf der rechten Seite von (4.49) maximal ist, wenn $\theta = \theta^{(0)}$ gilt, liegt auf der linken Seite von (4.49) genau dann ein Maximum vor, wenn der erste Term auf der rechten Seite dieser Gleichung maximal ist. Damit stellt jede Lösung von (4.51) ein Maximum von $f\left(A^{obs} \mid \theta\right)$ dar. Der Maximum-Likelihood-Schätzwert $\hat{\theta}$ erfüllt also die Gleichung (4.51) (vgl. *Beale, Little, 1975, S. 131-132*).

Die dargestellte Theorie soll nun zur Schätzung der Verteilungsparameter einer multivariat normalverteilten unvollständigen Stichprobe herangezogen werden, d.h. auf Basis einer unvollständigen, quantitativen Datenmatrix A, für deren m Merkmale eine m-variate Normalverteilung unterstellt wird, sind der Mittelwertvektor $\mu = (\mu_1, \ldots, \mu_m)$ und die Kovarianzmatrix $\Sigma = (\sigma_{kl})_{m,m}$ zu schätzen. In der obigen Notation ist

$$\theta = (\mu, \Sigma), \quad \theta^{(0)} = \left(\mu^{(0)}, \Sigma^{(0)}\right), \quad \theta^{*} = g\left(\theta^{(0)}\right) = \left(\mu^{*}, \Sigma^{*}\right) \tag{4.52}$$

und die logarithmierte Likelihood-Funktion besitzt die Form

$$\ln f\left(A \mid \mu, \Sigma\right) = -\tfrac{1}{2} \cdot n \cdot m \cdot \ln(2\pi) - \tfrac{1}{2} \cdot n \cdot \ln(\det \Sigma) - \tfrac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{m} \sum_{l=1}^{m} (a_{ik} - \mu_k) \sigma^{kl} (a_{il} - \mu_l). \tag{4.53}$$

Dabei stellt σ^{kl} das kl-te Element von Σ^{-1} dar. Der Erwartungswert der logarithmierten Likelihood-Funktion auf Basis der vorhandenen Daten A^{obs} sowie der Parameter $\mu^{(0)}$ und $\Sigma^{(0)}$ ergibt sich dann gemäß

$$E\left[\ln f\left(A \mid \mu, \Sigma\right) \mid A^{obs}, \mu^{(0)}, \Sigma^{(0)}\right] = -\frac{1}{2} \cdot n \cdot m \cdot \ln(2\pi) - \frac{1}{2} \cdot n \cdot \ln(\det \Sigma) \tag{4.54}$$

$$-\frac{1}{2} \cdot \sum_{i=1}^{n} \sum_{k=1}^{m} \sum_{l=1}^{m} \left[\left(E\left(a_{ik} \mid a^{i,obs}, \mu^{(0)}, \Sigma^{(0)}\right) - \mu_k \right) \sigma^{kl} \left(E\left(a_{il} \mid a^{i,obs}, \mu^{(0)}, \Sigma^{(0)}\right) - \mu_l \right) \right.$$

$$\left. + Cov\left(a_{ik}, a_{il} \mid a^{i,obs}, \mu^{(0)}, \Sigma^{(0)}\right) \right],$$

wobei $a^{i,obs}$ die vorhandenen Ausprägungen bei Objekt i andeutet. Eine Maximierung der in (4.54) angegebenen Likelihood-Funktion nach den Parametern μ und Σ führt zu dem (4.50) entsprechenden Ergebnis

$$\mu^{*} = \left(\mu_k^{*}\right)_{m,1}, \; \Sigma^{*} = \left(\sigma_{kl}^{*}\right)_{m,m} \quad \text{mit} \quad \mu_k^{*} = \frac{1}{n} \sum_{i=1}^{n} E\left(a_{ik} \mid a^{i,obs}, \mu^{(0)}, \Sigma^{(0)}\right), \tag{4.55}$$

$$\sigma_{kl}^{*} = \frac{1}{n} \sum_{i=1}^{n} \left[\left(E\left(a_{ik} \mid a^{i,obs}, \mu^{(0)}, \Sigma^{(0)}\right) - \mu_k^{*} \right) \left(E\left(a_{il} \mid a^{i,obs}, \mu^{(0)}, \Sigma^{(0)}\right) - \mu_l^{*} \right) \right.$$

$$\left. + Cov\left(a_{ik}, a_{il} \mid a^{i,obs}, \mu^{(0)}, \Sigma^{(0)}\right) \right].$$

Mit $\mu^{(0)} = \mu^* = \mu$ und $\Sigma^{(0)} = \Sigma^* = \Sigma$ ergeben sich dann analog zu (4.51) die folgenden Gleichungen:

$$A = (a_{ik})_{n,m} \quad \text{mit} \quad a_{ik} = \begin{cases} E\left(a_{ik} \middle| a^{i,obs}, \mu, \Sigma\right) & \text{falls } v_{ik} = 0 \\ a_{ik} & \text{falls } v_{ik} = 1 \end{cases}, \tag{4.56}$$

$$\mu = (\mu_k)_{m,1}, \Sigma = (\sigma_{kl})_{m,m} \quad \text{mit} \quad \mu_k = \frac{1}{n}\sum_{i=1}^{n} a_{ik}, \tag{4.57}$$

$$\sigma_{kl} = \frac{1}{n}\sum_{i=1}^{n}\left[(a_{ik} - \mu_k)(a_{il} - \mu_l) + c_{ikl}\right],$$

$$c_{ikl} = \begin{cases} Cov\left(a_{ik}, a_{il} \middle| a^{i,obs}, \mu, \Sigma\right) & \text{falls } v_{ik} = v_{il} = 0 \\ 0 & \text{sonst} \end{cases}.$$

Die Maximum-Likelihood-Schätzungen für den Mittelwertvektor und die Kovarianzmatrix resultieren danach aus einer Anfangsschätzung $\mu^{(0)}$ und $\Sigma^{(0)}$ für die Parameter sowie einer iterativen Berechnung der in (4.56) und (4.57) angegebenen Gleichungen. Der Iterationsprozeß wird solange durchgeführt, bis Konvergenz erreicht ist bzw. sich keine wesentlichen Veränderungen der geschätzten Parameter zwischen zwei aufeinanderfolgenden Iterationen ergeben, also ein entsprechend formuliertes Abbruchkriterium erfüllt ist. (4.56) stellt den sogenannten **E-Schritt** (expectation step) und (4.57) den sogenannten **M-Schritt** (maximization step) dar, woraus sich die Bezeichnung „EM-Algorithmus" ergibt.

Als Anfangsschätzung für die Parameter können beispielsweise der Mittelwertvektor und die Kovarianzmatrix der vollständig vorliegenden Objekte herangezogen werden. Beim E-Schritt werden dann Imputationswerte für die fehlenden Ausprägungen unter Verwendung der vorhandenen Daten und der Parameterschätzungen bestimmt. Dazu werden lineare Regressionsfunktionen, die auf Basis der geschätzten Parameter ermittelt werden, herangezogen. Der bedingte Erwartungswert für eine fehlende Ausprägung, der in (4.56) benötigt wird, kann somit gemäß der Formel

$$E\left(a_{ik} \middle| a^{i,obs}, \mu, \Sigma\right) = \mu_k + \Sigma_{k,M^{i,obs}} \cdot \Sigma^{-1}_{M^{i,obs},M^{i,obs}} \cdot \left(a^{i,obs} - \mu^{i,obs}\right) \tag{4.58}$$

bestimmt werden (vgl. *Schwab, 1991, S. 69*). Dabei bezeichnet $M^{i,obs}$ die Menge von Merkmalen, deren Ausprägungen beim Objekt i vorliegen. Entsprechend ist $\Sigma_{k,M^{i,obs}}$ die Teilmatrix von Σ, in der die Koeffizienten zwischen dem Merkmal k und den Merkmalen der Menge $M^{i,obs}$ enthalten sind. Diese Matrix besitzt damit eine Zeile und die Anzahl der Spalten entspricht der Anzahl der Merkmale in der Menge $M^{i,obs}$. Analog ist die Matrix $\Sigma_{M^{i,obs},M^{i,obs}}$ zu interpretieren. $a^{i,obs}$ ist der Vektor der vorhandenen Ausprägungen bei Objekt i und $\mu^{i,obs}$ der diesem Vektor entsprechende Mittelwertvektor. Aus der auf diese

Art vervollständigten Datenmatrix werden schließlich im M-Schritt die Mittelwerte, Varianzen und Kovarianzen erneut berechnet. Dabei werden alle Varianzen von Merkmalen mit fehlenden Daten sowie alle Kovarianzen von jeweils zwei Merkmalen mit fehlenden Daten mittels der entsprechenden, aus den multiplen Regressionsmodellen resultierenden Varianzen bzw. Kovarianzen der Residuen korrigiert. Der Anteil der für das betrachtete Merkmal jeweils fehlenden bzw. für die beiden betrachteten Merkmale jeweils gleichzeitig fehlenden Daten wird dabei berücksichtigt. Die Kovarianzen der Residuen, die in (4.57) bei dieser Korrektur heranzuziehen sind, ergeben sich gemäß

$$Cov\left(a_{ik}, a_{il} \,\middle|\, a^{i,obs}, \mu, \Sigma\right) = \gamma_{kl} \;\forall\, i = 1,\ldots,n \quad \text{mit} \quad \Gamma = \left(\gamma_{kl}\right)_{m,m} \tag{4.59}$$

$$\text{und} \quad \Gamma_{M^{mis},M^{obs}} = \Gamma_{M^{obs},M^{mis}} = \Gamma_{M^{obs},M^{obs}} = \begin{pmatrix} 0 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 0 \end{pmatrix},$$

$$\Gamma_{M^{mis},M^{mis}} = \Sigma_{M^{mis},M^{mis}} - \left(\Sigma_{M^{mis},M^{obs}} \cdot \Sigma^{-1}_{M^{obs},M^{obs}} \cdot \Sigma_{M^{obs},M^{mis}}\right)$$

(vgl. *Schwab, 1991, S. 71*). Dabei bezeichnet M^{obs} die Menge von Merkmalen, die vollständig vorliegen, und M^{mis} die Menge von Merkmalen, bei denen Daten fehlen. Entsprechend ist $\Gamma_{M^{mis},M^{obs}}$ die Teilmatrix von Γ, in der die Koeffizienten zwischen den Merkmalen der Menge M^{mis} (zeilenweise) und den Merkmalen der Menge M^{obs} (spaltenweise) enthalten sind. Analog sind die anderen Matrizen in (4.59) zu interpretieren.

Beispiel:

Für die Datenmatrix des Anhangs A (Fall 1) werden im folgenden lediglich die sechs kardinal skalierten Merkmale betrachtet, wobei eine 6-variate Normalverteilung unterstellt wird. Als Anfangsschätzungen für die Parameter werden der Mittelwertvektor und die Kovarianzmatrix herangezogen, die auf Basis der Objekte resultieren, die bei den sechs betrachteten Merkmalen vollständig vorliegen. Zur Bestimmung dieser nachfolgend dargestellten Anfangsschätzungen für die Parameter werden also die Objekte BMDP, CSS, MINITAB, NCSS, P-STAT, RS/1, SAS, SPSS und STATISTIX verwendet.

$$\mu^{(0)} = \begin{pmatrix} \mu_4^{(0)} \\ \vdots \\ \mu_9^{(0)} \end{pmatrix} = \begin{pmatrix} 78.37 \\ 74.33 \\ 40.22 \\ 54.26 \\ 52.56 \\ 40.25 \end{pmatrix}, \quad \Sigma^{(0)} = \begin{pmatrix} \sigma_{44}^{(0)} & \ldots & \sigma_{49}^{(0)} \\ \vdots & & \vdots \\ \sigma_{94}^{(0)} & \ldots & \sigma_{99}^{(0)} \end{pmatrix} = \begin{pmatrix} 179.98 & 43.39 & 227.23 & 192.51 & 176.86 & -44.54 \\ 43.39 & 162.00 & -55.38 & -11.94 & 86.37 & -109.77 \\ 227.23 & -55.38 & 501.39 & 428.46 & 572.34 & 154.95 \\ 192.51 & -11.94 & 428.46 & 567.08 & 801.63 & 342.92 \\ 176.86 & 86.37 & 572.34 & 801.63 & 1724.25 & 679.03 \\ -44.54 & -109.77 & 154.95 & 342.92 & 679.03 & 531.22 \end{pmatrix}.$$

Im E-Schritt der ersten Iteration ergeben sich die folgenden Imputationswerte für die fehlenden Daten, wobei die Berechnung am Beispiel der fehlenden Ausprägung $a_{12,4}$ dargestellt ist:

Fehlende Ausprägung	a_{24}	$a_{12,4}$	$a_{14,4}$	$a_{12,7}$	$a_{14,7}$	$a_{15,7}$	a_{49}	$a_{11,9}$
Imputationswert	45.18	58.67	47.30	37.14	5.89	97.09	2.62	74.24

$$a_{12,4} = \mu_4^{(0)} + \left(\sigma_{45}^{(0)}, \sigma_{46}^{(0)}, \sigma_{48}^{(0)}, \sigma_{49}^{(0)}\right) \begin{pmatrix} \sigma_{55}^{(0)} & \sigma_{56}^{(0)} & \sigma_{58}^{(0)} & \sigma_{59}^{(0)} \\ \sigma_{65}^{(0)} & \sigma_{66}^{(0)} & \sigma_{68}^{(0)} & \sigma_{69}^{(0)} \\ \sigma_{85}^{(0)} & \sigma_{86}^{(0)} & \sigma_{88}^{(0)} & \sigma_{89}^{(0)} \\ \sigma_{95}^{(0)} & \sigma_{96}^{(0)} & \sigma_{98}^{(0)} & \sigma_{99}^{(0)} \end{pmatrix}^{-1} \begin{pmatrix} a_{12,5} - \mu_5^{(0)} \\ a_{12,6} - \mu_6^{(0)} \\ a_{12,8} - \mu_8^{(0)} \\ a_{12,9} - \mu_9^{(0)} \end{pmatrix}$$

$$= 78.37 + (43.39, 227.23, 176.86, -44.54) \begin{pmatrix} 162.00 & -55.38 & 86.37 & -109.77 \\ -55.38 & 501.39 & 572.34 & 154.95 \\ 86.37 & 572.34 & 1724.25 & 679.03 \\ -109.77 & 154.95 & 679.03 & 531.22 \end{pmatrix}^{-1} \begin{pmatrix} 78.00 - 74.33 \\ 33.71 - 40.22 \\ 97.00 - 52.56 \\ 42.00 - 40.25 \end{pmatrix} = 58.67.$$

Im M-Schritt der ersten Iteration werden auf Basis der vervollständigten Datenmatrix anschließend neue Parameterschätzungen $\mu^{(1)}$ und $\Sigma^{(1)}$ ermittelt. Dabei wird die Berechnung des Teils der Residualkovarianzmatrix, der für die Korrektur der Kovarianzmatrix relevant ist, sowie beispielhaft die Bestimmung der Parameter $\mu_4^{(1)}$, $\sigma_{44}^{(1)}$, $\sigma_{45}^{(1)}$ und $\sigma_{47}^{(1)}$ gezeigt:

$$\mu^{(1)} = \begin{pmatrix} \mu_4^{(1)} \\ \vdots \\ \mu_9^{(1)} \end{pmatrix} = \begin{pmatrix} 71.36 \\ 72.53 \\ 35.45 \\ 44.92 \\ 54.07 \\ 39.92 \end{pmatrix}, \quad \Sigma^{(1)} = \begin{pmatrix} \sigma_{44}^{(1)} & \cdots & \sigma_{49}^{(1)} \\ \vdots & & \vdots \\ \sigma_{94}^{(1)} & \cdots & \sigma_{99}^{(1)} \end{pmatrix} = \begin{pmatrix} 265.12 & 43.47 & 263.57 & 393.46 & 197.99 & 110.21 \\ 43.47 & 118.92 & -30.72 & -3.79 & 39.36 & -91.21 \\ 263.57 & -30.72 & 439.33 & 539.68 & 509.13 & 264.96 \\ 393.46 & -3.79 & 539.68 & 977.15 & 849.61 & 598.53 \\ 197.99 & 39.36 & 509.13 & 849.61 & 1601.93 & 811.92 \\ 110.21 & -91.21 & 264.96 & 598.53 & 811.92 & 730.05 \end{pmatrix},$$

$$\Gamma_{M^{mis},M^{mis}} = \begin{pmatrix} \gamma_{44} & \gamma_{47} & \gamma_{49} \\ \gamma_{74} & \gamma_{77} & \gamma_{79} \\ \gamma_{94} & \gamma_{97} & \gamma_{99} \end{pmatrix} = \begin{pmatrix} 179.98 & 192.51 & -44.54 \\ 192.51 & 567.08 & 342.92 \\ -44.54 & 342.92 & 531.22 \end{pmatrix} - \left[\begin{pmatrix} 43.39 & 227.23 & 176.86 \\ -11.94 & 428.46 & 801.63 \\ -109.77 & 154.95 & 679.03 \end{pmatrix} \cdot \begin{pmatrix} 162.00 & -55.38 & 86.37 \\ -55.38 & 501.39 & 572.34 \\ 86.37 & 572.34 & 1724.25 \end{pmatrix}^{-1} \begin{pmatrix} 43.39 & -11.94 & -109.77 \\ 227.23 & 428.46 & 154.95 \\ 176.86 & 801.63 & 679.03 \end{pmatrix} \right] = \begin{pmatrix} 43.09 & 72.89 & 74.02 \\ 72.89 & 218.19 & 178.36 \\ 74.02 & 178.36 & 184.76 \end{pmatrix},$$

$$\mu_4^{(1)} = \tfrac{1}{15} \cdot (92.66 + 45.18 + 80.33 + 61.33 + \ldots + 65.66 + 58.67 + 67.33 + 47.30 + 86.66) = 71.36,$$

$$\begin{aligned} \sigma_{44}^{(1)} = \tfrac{1}{15} \cdot \big(& (92.66 - 71.36)^2 + (45.18 - 71.36)^2 + 43.09 + (80.33 - 71.36)^2 + (61.66 - 71.36)^2 + (62.33 - 71.36)^2 \\ & + (76.66 - 71.36)^2 + (92.66 - 71.36)^2 + (54.33 - 71.36)^2 + (89.66 - 71.36)^2 + (89.33 - 71.36)^2 \\ & + (65.66 - 71.36)^2 + (58.67 - 71.36)^2 + 43.09 + (67.33 - 71.36)^2 + (47.30 - 71.36)^2 + 43.09 \\ & + (86.66 - 71.36)^2 \big) = 265.12, \end{aligned}$$

$$\sigma_{45}^{(1)} = \tfrac{1}{15} \cdot \big((92.66 - 71.36)(56.00 - 72.53) + \ldots + (86.66 - 71.36)(64.00 - 72.53) \big) = 43.47,$$

$$\begin{aligned} \sigma_{47}^{(1)} = \tfrac{1}{15} \cdot \big(& (92.66 - 71.36)(48.66 - 44.92) + (45.18 - 71.36)(0.00 - 44.92) + (80.33 - 71.36)(77.00 - 44.92) \\ & + (61.66 - 71.36)(2.66 - 44.92) + (62.33 - 71.36)(31.00 - 44.92) + (76.66 - 71.36)(80.33 - 44.92) \\ & + (92.66 - 71.36)(55.66 - 44.92) + (54.33 - 71.36)(27.66 - 44.92) + (89.66 - 71.36)(94.00 - 44.92) \\ & + (89.33 - 71.36)(52.00 - 44.92) + (65.66 - 71.36)(42.66 - 44.92) + (58.67 - 71.36)(37.14 - 44.92) \\ & + 72.89 + (67.33 - 71.36)(22.00 - 44.92) + (47.30 - 71.36)(5.89 - 44.92) + 72.89 \\ & + (86.66 - 71.36)(97.09 - 44.92) \big) = 393.46. \end{aligned}$$

Ausgehend von den am Ende einer Iteration geschätzten Parametern werden der E- und der M-Schritt solange wiederholt, bis die Abweichungen aller geschätzten Parameter in zwei aufeinanderfolgenden Itera-

tionen kleiner als 0.0001 sind. Dieses Abbruchkriterium ist in diesem Beispiel nach 47 Iterationen erfüllt. Es resultieren die folgenden, endgültigen Schätzungen für den Mittelwertvektor und die Kovarianzmatrix:

$$\mu^* = \begin{pmatrix} 72.76 \\ 72.53 \\ 35.45 \\ 45.31 \\ 54.07 \\ 38.64 \end{pmatrix}, \quad \Sigma^* = \begin{pmatrix} 210.97 & 39.16 & 245.33 & 332.41 & 197.63 & 100.91 \\ 39.16 & 118.92 & -30.72 & -0.83 & 39.36 & -91.98 \\ 245.33 & -30.72 & 439.33 & 522.72 & 509.13 & 283.18 \\ 332.41 & -0.83 & 522.72 & 901.63 & 846.91 & 602.86 \\ 197.63 & 39.36 & 509.13 & 846.91 & 1601.93 & 820.39 \\ 100.91 & -91.98 & 283.18 & 602.86 & 820.39 & 738.66 \end{pmatrix}.$$

Falls ein einziger, endlicher Maximum-Likelihood-Schätzwert der Parameter existiert, wird dieser Schätzwert mit dem EM-Algorithmus auch gefunden. Die Konvergenz des Algorithmus ist auf jeden Fall gesichert, sobald die Likelihood-Funktion beschränkt ist. Wie das vorhergehende Beispiel verdeutlicht, ist die Konvergenzgeschwindigkeit des EM-Algorithmus jedoch verhältnismäßig gering. Die Konvergenzgeschwindigkeit hängt dabei vom Anteil der fehlenden Daten ab. Je höher der Anteil der fehlenden Daten ist, desto geringer ist die Konvergenzgeschwindigkeit. Eine ausführliche Darstellung der Konvergenzeigenschaften sowie der Konvergenzgeschwindigkeit des EM-Algorithmus ist bei *Dempster et al. (1977, S. 6-11)* zu finden. Dabei ist anzumerken, daß der bei *Dempster et al.* angegebene Konvergenzbeweis laut *Laird (1985, S. 550)* inkorrekt sein soll.

4.3.3 Vergleich und Kombination der Verfahren

Die hier in erster Linie betrachteten Verfahren zur Schätzung von Mittelwerten, Varianzen und Kovarianzen unterscheiden sich vor allem hinsichtlich der jeweils zugrundeliegenden Theorie. Neben den größtenteils benötigten Verteilungsannahmen wird die Anwendung einer Reihe von Verfahren auch durch die jeweils vorausgesetzten, speziellen Muster fehlender Daten eingeschränkt. Sofern nicht ein systematischer Ausfallmechanismus vorliegt, ist lediglich der EM-Algorithmus bzw. die dazu identische, iterative Anwendung der Methode von *Buck* ein grundsätzlich geeignetes Verfahren. In der Tabelle 4.3 sind die Voraussetzungen sowie die wichtigsten Eigenschaften der betrachteten Parameterschätzverfahren für stetige Daten noch einmal zusammenfassend dargestellt.

Abschließend ist noch auf einen Ansatz von *Little und Schluchter (1985)* hinzuweisen, bei dem Parameterschätzverfahren für kategoriale und stetige Daten kombiniert werden. Für den Fall einer unvollständigen Datenmatrix mit kardinalen und nominalen Merkmalen wird dabei ein Modell konstruiert, auf dessen Grundlage eine Maximum-Likelihood-Schätzung für die erwarteten Zellhäufigkeiten bei den nominalen Merkmalen sowie für die in den einzelnen Zellen sich ergebenden Mittelwerte, Varianzen und Kovarianzen der kardinalen Merkmale durchgeführt wird.

Verfahren	Voraussetzungen	Eigenschaften
Maximum-Likelihood-Schätzung nach *Wilks*	Datenmatrix (zwei Merkmale), bivariate Normalverteilung, Daten sind MAR	Explizite Lösung der Maximum-Likelihood-Gleichungen für die Verteilungsparameter ist nicht möglich
Maximum-Likelihood-Schätzung nach *Matthai*	Datenmatrix (drei Merkmale), trivariate Normalverteilung, Daten sind MAR	Explizite Lösung der Maximum-Likelihood-Gleichungen für die Verteilungsparameter ist nicht möglich
Maximum-Likelihood-Schätzung nach *Lord*	Datenmatrix (drei Merkmale, MD in zwei Merkmalen mit speziellem Muster), trivariate Normalverteilung, Daten sind MAR	Explizite Lösung der Maximum-Likelihood-Gleichungen für die Verteilungsparameter ist für das vorausgesetzte Muster der fehlenden Daten möglich
Maximum-Likelihood-Schätzung nach *Edgett*	Datenmatrix (drei Merkmale, MD in einem Merkmal), trivariate Normalverteilung, Daten sind MAR	Explizite Lösung der Maximum-Likelihood-Gleichungen für die Verteilungsparameter ist für das vorausgesetzte Muster der fehlenden Daten möglich
Maximum-Likelihood-Schätzung durch Faktorisierung der Likelihood	Datenmatrix, multivariate Normalverteilung, nested MD-Muster, Daten sind MAR	Explizite Lösung der Maximum-Likelihood-Gleichungen für die Verteilungsparameter ist für das vorausgesetzte Muster der fehlenden Daten möglich
EM-Algorithmus	Datenmatrix, multivariate Normalverteilung, Daten sind MAR	Iterative Lösung der Maximum-Likelihood-Gleichungen für die Verteilungsparameter bei beliebigen MD-Mustern
Maximum-Likelihood-Schätzung auf Basis eines Ausfallmodells	Datenmatrix, multivariate Normalverteilung, Modell des Ausfallmechanismus	Ausfallmodell wird in der Maximum-Likelihood-Funktion berücksichtigt, Schätzung der Verteilungsparameter mit den Verfahren auf Basis eines unsystematischen Ausfallmechanismus
Bayes-Schätzung	Datenmatrix, multivariate Normalverteilung, Daten sind MAR	Vorinformationen der zu schätzenden Verteilungsparameter auf Basis des Stichprobenmechanismus werden benötigt und zur Schätzung herangezogen
Bayes-Schätzung auf Basis eines Ausfallmodells	Datenmatrix, multivariate Normalverteilung, Modell des Ausfallmechanismus	Stichprobenmechanismus muß bekannt sein und wird mit dem Ausfallmechanismus zur Schätzung der Verteilungsparameter herangezogen
Schätzung nach *Buck*	Datenmatrix (quantitativ), Daten sind MCAR oder MAR	Verhältnismäßig einfache Anwendbarkeit, Ergebnis ist eine unverzerrte Schätzung für die Kovarianzmatrix
Gewichtungsmethoden	Datenmatrix (quantitativ), Daten sind MCAR oder MCARC	Stichprobenmechanismus muß bekannt sein, Ergebnis ist eine unverzerrte Schätzung für die Mittelwerte

Tabelle 4.3: Vergleich der Parameterschätzverfahren für stetige Daten

4.4 Multivariate Analyseverfahren

Die in diesem Abschnitt betrachteten multivariaten Analyseverfahren stellen MD-Verfahren dar, die ausgehend von einer unvollständigen Daten- bzw. Distanzmatrix unmittelbar zu den jeweiligen Analyseergebnissen führen. Diese Verfahren verwenden ausschließlich die vorliegenden Daten und können die fehlenden Werte entsprechend berücksichtigen. Dies wird durch im allgemeinen geringfügige Modifikationen der jeweiligen, auf vollständigen Daten basierenden Auswertungsmethoden erreicht, d.h. die verfahrensspezifischen Möglichkeiten werden ausgenutzt, so daß fehlende Daten zugelassen sind.

Damit ergibt sich eine Abgrenzung zu den bereits vorgestellten MD-Verfahren bzw. Strategien: Die hier betrachteten multivariaten Analyseverfahren basieren nicht auf zuvor durchgeführten Parameterschätzungen, verwenden keine Imputationswerte für die fehlenden Daten und beruhen nicht auf einer Eliminierungsstrategie. So zählt beispielsweise der **Ansatz von *Wishart*** *(1978, 1985, 1986)* nicht zu den MD-Verfahren dieses Abschnitts, da *Wishart* die Clusteranalysemethoden Ward, Median, Average Linkage und k-Means auf Basis eines Klassifikations- bzw. Verschiedenheitsindexes durchführt, bei dem die für zwei Objekte jeweils verfügbaren Merkmale herangezogen werden. Diesem Ansatz liegt somit ein Eliminierungsverfahren zugrunde.

Eine Einteilung der nachfolgend dargestellten multivariaten Analyseverfahren erfolgt im Hinblick auf die jeweilige Analysemethode. Dabei wird in **Abschnitt 4.4.1** die Clusteranalyse, in **Abschnitt 4.4.2** die multidimensionale Skalierung, in **Abschnitt 4.4.3** die Faktorenanalyse und in **Abschnitt 4.4.4** die Regressionsanalyse betrachtet. In **Abschnitt 4.4.5** erfolgt abschließend ein Vergleich der vorgestellten Verfahren. Da diese Verfahren unmittelbar auf der unvollständigen Daten- oder Distanzmatrix basieren, müssen die Daten grundsätzlich der Bedingung MCAR genügen. Lediglich bei Berücksichtigung aller Abhängigkeitsbeziehungen der fehlenden Daten von den vorhandenen Daten reicht die weniger restriktive MAR-Annahme aus.

4.4.1 Clusteranalyse

Für agglomerative hierarchische Klassifikationsverfahren wird von *Schader und Gaul (1991)* eine geringfügige Modifikation des grundsätzlichen Verfahrensprinzips und der entsprechenden Verfahrensvarianten zur Berücksichtigung fehlender Daten vorgeschlagen. Dieses Verfahren wird als **MVL Verfahren** (missing value linkage) bezeichnet und basiert auf einer unvollständigen, unmittelbar erhobenen Distanzmatrix. Das Verfah-

rensprinzip des MVL Ansatzes ist in der Abbildung 4.5 dargestellt, wobei $v(K,L)$ einen Verschiedenheitsindex zwischen den Klassen K und L bezeichnet.

Objektmenge N, Index $v(K,L)$, $\mathcal{K}^v = \{K_1^v,\ldots,K_{n-v}^v\}$, $v = 0$

wiederhole bis $|\mathcal{K}^v| = 1$ oder $v(K,L) =$ fehlend $\forall\, K,L \in \mathcal{K}^v$

Suche Klassen mit minimaler Verschiedenheit

$$\min_{\substack{K_i^v,K_j^v \in \mathcal{K}^v \\ K_i^v \neq K_j^v}} v\left(K_i^v,K_j^v\right) = v\left(K_{i_0}^v,K_{j_0}^v\right) \neq \text{fehlend}$$

Bestimme Klassifikation $\mathcal{K}^{v+1} = \{K_1^{v+1},\ldots,K_{n-v-1}^{v+1}\}$

$$\text{mit } K_i^{v+1} = \begin{cases} K_{i_0}^v \cup K_{j_0}^v & \text{für } i = \min\{i_0,j_0\} \\ K_{i+1}^v & \text{für } i \geq \max\{i_0,j_0\} \\ K_i^v & \text{sonst} \end{cases}$$

Berechne neue Verschiedenheiten $v\left(K_i^{v+1},K_j^{v+1}\right)$

$v = v+1$

Abbildung 4.5: Verfahrensprinzip des MVL Ansatzes

Gegenüber dem Verfahrensprinzip bei Vorliegen einer vollständigen Distanzmatrix (vgl. z.B. *Opitz, 1980, S. 96-97*) unterscheidet sich das Verfahrensprinzip des MVL Ansatzes in zwei Punkten. Zum einen können die Klassen mit minimaler Verschiedenheit lediglich auf Basis der vorhandenen Verschiedenheitsindizes ausgewählt werden. Zum anderen wird als Abbruchkriterium bei der iterativen Bestimmung von immer gröberen Zerlegungen der Objektmenge nicht nur das Vorliegen der gröbsten Zerlegung in Form einer Klasse mit allen Objekten herangezogen, sondern das Verfahren wird auch dann abgebrochen, wenn alle Verschiedenheitsindizes zwischen den Klassen der zuletzt bestimmten Zerlegung fehlende Werte darstellen.

In der Tabelle 4.4 ist für die relevanten Verfahrensvarianten angegeben, wie die neuen Verschiedenheiten unmittelbar zu bestimmen sind. Bei dieser Berechnung zeigt sich dann, welche der zunächst noch fehlenden Verschiedenheitsindizes bei einer Fusion von

zwei Klassen wegfallen. Dabei sind in der Menge $(N \times N)^{mis}$ alle Objektpaare zusammengefaßt, deren paarweise Distanzen fehlen, d.h. $(N \times N)^{mis} = \{(i,j): w_{ij} = 0\}$.

Verfahrensvariante	**Verschiedenheitsindex**
Single Linkage	$v(K,L) = \begin{cases} \min\{d_{ij}: (i,j) \in (K \times L) - (N \times N)^{mis}\} & \text{falls } (K \times L) - (N \times N)^{mis} \neq \emptyset \\ \text{fehlend} & \text{sonst} \end{cases}$
Complete Linkage	$v(K,L) = \begin{cases} \max\{d_{ij}: (i,j) \in (K \times L) - (N \times N)^{mis}\} & \text{falls } (K \times L) - (N \times N)^{mis} \neq \emptyset \\ \text{fehlend} & \text{sonst} \end{cases}$
Average Linkage	$v(K,L) = \begin{cases} \frac{1}{\left\lvert (K \times L) - (N \times N)^{mis} \right\rvert} \cdot \sum_{(i,j) \in (K \times L) - (N \times N)^{mis}} d_{ij} & \text{falls } (K \times L) - (N \times N)^{mis} \neq \emptyset \\ \text{fehlend} & \text{sonst} \end{cases}$

Tabelle 4.4: Verschiedenheitsindizes von Verfahrensvarianten des MVL Ansatzes

Die Verschiedenheit zwischen zwei Klassen wird danach aus den vorhandenen Distanzen zwischen den Objekten der beiden Klassen bestimmt bzw., für den Fall keiner einzigen vorhandenen Distanz, als fehlend betrachtet. Neben der unmittelbaren Bestimmung der Verschiedenheitsindizes für die einzelnen Verfahrensvarianten ist auch die rekursive Berechnung der Zwischenklassenverschiedenheiten gemäß der Formel nach *Lance und Williams*, die im Hinblick auf die fehlenden Daten entsprechend zu modifizieren ist, möglich (vgl. *Schader, Gaul, 1991, S. 109*). Damit ist auch eine Anwendung der Verfahren Flexible Strategy, Median, Centroid und Ward auf Basis der Distanzmatrix denkbar. Dabei ist allerdings zu beachten, daß die Verfahren Median, Centroid und Ward im allgemeinen bei Vorliegen einer quantitativen Datenmatrix, aus der die Verschiedenheiten zwischen den Klassen zu bestimmen sind, angewandt werden (vgl. *Opitz, 1980, S. 99*).

Die aus dem MVL Verfahren resultierende Hierarchie kann entweder vollständig oder unvollständig sein. Eine vollständige Hierarchie liegt dann vor, wenn das Verfahren erst nach der Bestimmung der gröbsten Zerlegung in Form einer Klasse mit allen Objekten abgebrochen wird. Der Wegfall fehlender Verschiedenheitsindizes im Laufe des Fusionsprozesses hängt dabei vom Anteil und den Konzentrationstendenzen der fehlenden Distanzen ab. Bei einem niedrigen Anteil sowie einer gleichmäßigen Streuung der fehlenden Distanzen über die Matrix ist ein vollständiger Wegfall der fehlenden Verschiedenheitsindizes im Laufe des Fusionsprozesses eher zu erwarten als bei einem hohen Anteil sowie einem konzentrierten Auftreten der fehlenden Distanzen in einem Bereich der Matrix.

Beispiel:

Für die Distanzmatrix des Anhangs B (Fall 1) soll im folgenden mittels des MVL Ansatzes die Complete Linkage Lösung bestimmt werden. Unter Verwendung der Objektnummern, die gemäß der Anordnung der Objekte in der Distanzmatrix resultieren, ergeben sich die folgenden Rechenschritte des Fusionsprozesses, wobei jeweils die obere Dreiecksmatrix der Verschiedenheitsindizes angegeben und das Minimum dieser Werte umrahmt ist:

	2	3	4	5	6	7	8	9	10
1	2.57	7.88		4.95	3.95	4.25	4.65	5.23	
2		9.43	5.32	5.33	4.45		4.76	5.74	
3			5.65		5.48		3.55	4.32	6.35
4				2.35		2.65	3.45	2.08	4.28
5					4.96	2.06	3.58		4.12
6						3.02	3.85		3.25
7							2.05	7.18	1.92
8								5.23	1.78
9									6.56

	2	3	4	5	6	7	8,10	9
1	2.57	7.88		4.95	3.95	4.25	4.65	5.23
2		9.43	5.32	5.33	4.45		4.76	5.74
3			5.65		5.48		6.35	4.32
4				2.35		2.65	4.28	2.08
5					4.96	2.06	4.12	
6						3.02	3.85	
7							2.05	7.18
8,10								6.56

	2	3	4	5	6	7,8,10	9
1	2.57	7.88		4.95	3.95	4.65	5.23
2		9.43	5.32	5.33	4.45	4.76	5.74
3			5.65		5.48	6.35	4.32
4				2.35		4.28	2.08
5					4.96	4.12	
6						3.85	
7,8,10							7.18

	2	3	4,9	5	6	7,8,10
1	2.57	7.88	5.23	4.95	3.95	4.65
2		9.43	5.74	5.33	4.45	4.76
3			5.65		5.48	6.35
4,9				2.35		7.18
5					4.96	4.12
6						3.85

	2	3	4,5,9	6	7,8,10
1	2.57	7.88	5.23	3.95	4.65
2		9.43	5.74	4.45	4.76
3			5.65	5.48	6.35
4,5,9				4.96	7.18
6					3.85

	3	4,5,9	6	7,8,10
1,2	9.43	5.74	4.45	4.76
3		5.65	5.48	6.35
4,5,9			4.96	7.18
6				3.85

	3	4,5,9	6,7,8,10
1,2	9.43	5.74	4.76
3		5.65	6.35
4,5,9			7.18

	3	4,5,9
1,2,6,7,8,10	9.43	7.18
3		5.65

	3,4,5,9
1,2,6,7,8,10	9.43

In diesem Beispiel resultiert eine vollständige Hierarchie, da alle fehlenden Verschiedenheitsindizes im Laufe des Fusionsprozesses wegfallen. Dieses Ergebnis war aufgrund des zufälligen Fehlens und der dadurch bedingten gleichmäßigen Streuung der paarweisen Distanzen über die Matrix sowie des geringen

Anteils an fehlenden Daten auch zu erwarten. In der Abbildung 4.6 ist die Hierarchie noch einmal in Form eines Dendrogramms dargestellt.

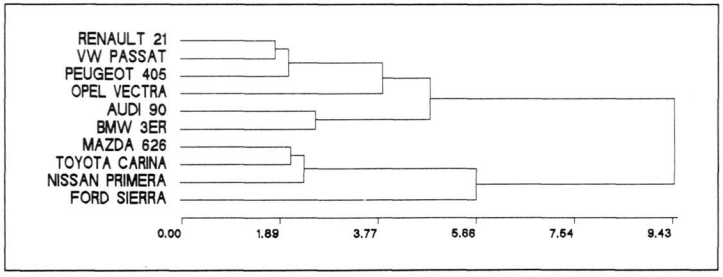

Abbildung 4.6: Complete Linkage Lösung für die 10 Mittelklasseautomobile

In einer weiteren Arbeit von *Schader und Gaul (1990)* wird die Modifikation eines pyramidalen Klassifikationsverfahrens zur Berücksichtigung fehlender Daten vorgestellt. Auf Basis einer unvollständigen, unmittelbar erhobenen Distanzmatrix wird der sogenannte **PAC Algorithmus** (pyramidal ascending clustering) analog zum MVL Verfahren modifiziert. Dabei wird die Verschiedenheit zwischen zwei Klassen aus den vorhandenen Distanzen zwischen den Objekten der beiden Klassen bestimmt bzw., für den Fall keiner einzigen vorhandenen Distanz, als fehlend betrachtet. Der Fusionsprozeß ist solange durchzuführen, wie Verschiedenheiten zwischen den Klassen vorhanden sind. Da der von *Schader und Gaul* benannte **PACII Algorithmus** (pyramidal ascending clustering with incomplete information) auf derselben Idee wie das MVL Verfahren basiert, wird an dieser Stelle auf eine ausführliche Darstellung dieses Ansatzes verzichtet.

4.4.2 Multidimensionale Skalierung

Die multidimensionale Skalierung stellt sich die Aufgabe, eine Anordnung der Objekte in einem Raum möglichst niedriger Dimension so zu bestimmen, daß die relative Lage der sich ergebenden Punkte die Ähnlichkeit bzw. Unähnlichkeit der Objekte angemessen beschreibt. Der Ausgangspunkt für die Verfahren der multidimensionalen Skalierung ist im allgemeinen eine Distanzmatrix. Da in einer vollständigen Distanzmatrix ein Teil der paarweisen Distanzen eine redundante Information darstellt, ist grundsätzlich auch im Fall einer unvollständigen Distanzmatrix eine adäquate Anordnung der Objekte möglich. Darüber hinaus können in einer unmittelbar erhobenen Distanzmatrix widersprüchliche Distanzen vorliegen, so daß selbst bei vollständig vorhandenen Distanzen gewisse Verzerrungen der Ergebnisse auftreten können (vgl. *Spence, Domoney, 1974, S. 469*).

Falls in einer unvollständigen Distanzmatrix genau die Distanzen fehlen, die eine redundante Information darstellen, dann kann eine angemessene Anordnung der Objekte auf Basis der vorhandenen Distanzen erfolgen. Wie *Spence und Domoney (1974, S. 478-479, 485)* sowie *Malhotra et al. (1988, S. 99)* in ihrer Simulationsstudie jedoch zeigen, führen zufällig fehlende Distanzen zu in etwa gleich guten Resultaten, sofern der Anteil der fehlenden Distanzen gewisse Grenzen nicht übersteigt. Nach *Spence und Domoney (1974, S. 484-485, 487)* können beispielsweise bei einer Anzahl von 32 Objekten bis zu 50 Prozent und bei einer Anzahl von 48 Objekten bis zu 70 Prozent der Distanzen fehlen, um dennoch eine angemessene Anordnung der Objekte im Hinblick auf deren Ähnlichkeit zu erhalten. Bei den von *Spence und Domoney* betrachteten 32 Objekten resultieren bei einem Anteil von bis zu ca. 30 Prozent fehlender Distanzen fast die gleichen Ergebnisse wie im Fall der vollständigen Distanzmatrix. Grundsätzlich sind die Ergebnisse um so besser, je geringer der Anteil der fehlenden Distanzen ist (vgl. *Malhotra et al., 1988, S. 97*). Darüber hinaus ist bei Vorliegen fehlender Distanzen das unter Umständen auftretende Problem zu berücksichtigen, daß die relative Lage eines Objektes zu den anderen Objekten aufgrund einer zu geringen Anzahl vorhandener bzw. keiner einzigen vorhandenen Distanz nicht eindeutig bzw. gar nicht bestimmt werden kann. Im Fall zufällig fehlender Distanzen ist im allgemeinen jedoch gewährleistet, daß der Anteil der fehlenden Distanzen für die einzelnen Objekte in etwa gleich groß ist.

Die meisten nonmetrischen Algorithmen der multidimensionalen Skalierung, unter anderem auch das **Verfahren von Kruskal**, können fehlende Daten berücksichtigen und damit auf Basis der vorhandenen Distanzen eine Anordnung der Objekte bestimmen. Diese Tatsache stellt einen Vorteil der nonmetrischen gegenüber den metrischen Verfahren dar *(Spence, Domoney, 1974, S.469-470)*. Betrachtet man beispielsweise das Verfahren von Kruskal, dann ist im Fall fehlender Distanzen eine Modifikation des Algorithmus dahingehend durchzuführen, daß bei der Bestimmung des Stresswerts einer Konfiguration sowie bei der Verbesserung dieser Konfiguration mittels des Gradientenverfahrens lediglich die vorhandenen empirischen sowie die entsprechenden, auf Basis der Konfiguration bestimmten Distanzen verwendet werden. Da die im Fall einer unvollständigen Distanzmatrix notwendige Modifikation des Verfahren verhältnismäßig geringfügig ist, wird an dieser Stelle auf eine ausführliche Darstellung verzichtet. Die Vorgehensweise soll lediglich durch das nachfolgende Beispiel verdeutlicht werden. Eine formale Beschreibung des Verfahrens von Kruskal bei Vorliegen einer vollständigen Distanzmatrix kann beispielsweise *Opitz (1980, S. 130-143)* entnommen werden.

Beispiel:

Für die Distanzmatrix des Anhangs B (Fall 1) werden zur Vereinfachung dieses Beispiels im folgenden lediglich die ersten fünf Objekte Audi 90 (1), BMW 3er (2), Ford Sierra (3), Mazda 626 (4) und Nissan Primera (5) betrachtet. Mit dem Verfahren von Kruskal soll eine Repräsentation dieser Objekte im $\mathbb{R}^1$ durch-

geführt werden. Dabei wird von der zufällig gewählten Startkonfiguration $X^{(0)T} = (x_{ik}^{(0)})_{5,1}^{T} = (0,2,5,2,2)$ ausgegangen. Zur Berechnung des Stresswerts für die Startkonfiguration erfolgt die nachfolgend dargestellte monotone Anpassung nach Kruskal auf Basis der vorhandenen paarweisen Distanzen. Im einzelnen bezeichnen d_{ij} die empirischen Distanzen, $\hat{d}_{ij}^{(0)}$ die L_1-Distanzen der Startkonfiguration und $\delta_{ij}^{(0)}$ die der Monotoniebedingung bei minimalem Stress genügenden Distanzen.

(i,j)	(4,5)	(1,2)	(1,5)	(2,4)	(2,5)	(3,4)	(1,3)	(2,3)
d_{ij}	2.35	2.57	4.95	5.32	5.33	5.65	7.88	9.43
$\hat{d}_{ij}^{(0)}$	0	2	2	0	0	3	5	3
$\delta_{ij}^{(0)}$	0	1	1	1	1	3	4	4

Damit ergeben sich die folgenden Stresswerte $b_0\left(X^{(0)}\right)$ (Rohstress), $b_{max}\left(X^{(0)}\right)$ (maximaler Stress) und $b_{norm}\left(X^{(0)}\right)$ (normierter Stress) für die Startkonfiguration $X^{(0)}$:

$$b_0\left(X^{(0)}\right) = \sum_{\substack{i,j:\, w_{ij}=1 \\ i<j}} \left(\hat{d}_{ij}^{(0)} - \delta_{ij}^{(0)}\right)^2 = 6, \quad b_{max}\left(X^{(0)}\right) = \sum_{\substack{i,j:\, w_{ij}=1 \\ i<j}} \left(\hat{d}_{ij}^{(0)} - \bar{d}^{(0)}\right)^2 = 22.875,$$

$$b_{norm}\left(X^{(0)}\right) = \frac{b_0\left(X^{(0)}\right)}{b_{max}\left(X^{(0)}\right)} \approx 0.26 \quad \text{mit} \quad \bar{d}^{(0)} = \frac{1}{w^{obs}} \sum_{\substack{i,j:\, w_{ij}=1 \\ i<j}} \left(\hat{d}_{ij}^{(0)}\right)^2 = \frac{1}{8} \cdot 15 = 1.875.$$

Da der Stresswert für die Startkonfiguration noch nicht zufriedenstellend ist, erfolgt eine Verbesserung dieser Konfiguration unter Anwendung des Gradientenverfahrens. In diesem Beispiel ergibt sich der Gradient der Stressfunktion gemäß

$$B^{(0)T} = \left(\frac{\partial b_0}{\partial x_{ik}}\right)_{5,1}^{T}\Bigg|_{X^{(0)}} = (-6,4,0,0,2), \quad \text{wobei} \quad \frac{\partial b_0}{\partial x_{ik}}\Bigg|_{X^{(0)}} = \sum_{\substack{j:\, w_{ij}=1 \\ j \neq i}} 2 \cdot \left(\hat{d}_{ij}^{(0)} - \delta_{ij}^{(0)}\right) \cdot \operatorname{sgn}\left(x_{ik}^{(0)} - x_{jk}^{(0)}\right).$$

Unter Verwendung dieses Gradienten und einer gewählten Schrittweite $\lambda_0 = 0.2$ kann schließlich die Konfiguration $X^{(1)}$ folgendermaßen bestimmt werden:

$$X^{(1)} = X^{(0)} - \lambda_0 \cdot B^{(0)} = \begin{pmatrix} 0 \\ 2 \\ 5 \\ 2 \\ 2 \end{pmatrix} - 0.2 \cdot \begin{pmatrix} -6 \\ 4 \\ 0 \\ 0 \\ 2 \end{pmatrix} = \begin{pmatrix} 1.2 \\ 1.2 \\ 5 \\ 2 \\ 1.6 \end{pmatrix}.$$

Die monotone Anpassung auf Basis der Konfiguration $X^{(1)}$ sowie die Bestimmung der entsprechenden Stresswerte führen zu folgendem Ergebnis:

(i,j)	(4,5)	(1,2)	(1,5)	(2,4)	(2,5)	(3,4)	(1,3)	(2,3)
d_{ij}	2.35	2.57	4.95	5.32	5.33	5.65	7.88	9.43
$\hat{d}_{ij}^{(1)}$	0.4	0	0.4	0.8	0.4	3	3.8	3.8
$\delta_{ij}^{(1)}$	0.2	0.2	0.4	0.6	0.6	3	3.8	3.8

$b_0\left(X^{(1)}\right) = 0.16, \quad b_{max}\left(X^{(1)}\right) = 19.155, \quad b_{norm}\left(X^{(1)}\right) \approx 0.008.$

Da der normierte Stresswert von 0.008 als sehr gut bezeichnet werden kann, wird das Verfahren abgebrochen. Die Konfiguration $X^{(1)}$ stellt damit eine Anordnung der Objekte im $\mathbb{R}^1$ dar, mit der die Ähnlichkeitsbeziehungen der Objekte auf Basis der vorhandenen Distanzen angemessen zum Ausdruck gebracht werden.

Bei der Anwendung des Gradientenverfahrens zur Verbesserung der Konfiguration ist eine Modifikation der Schrittweite in der Art denkbar, daß für jedes einzelne Objekt die Schrittweite explizit unter Berücksichtigung des Anteils der fehlenden Distanzen bei diesem Objekt gewählt wird. Ein Objekt mit wenigen vorhandenen Distanzen kann im Hinblick auf die Einhaltung der Monotoniebedingung problemloser verschoben werden als ein Objekt mit vielen vorhandenen Distanzen. Daher sollte die objektweise Schrittweite bei einem hohen Anteil fehlender Distanzen des jeweiligen Objekts verhältnismäßig groß und bei einem niedrigen Anteil entsprechend klein gewählt werden. Selbst bei geringen Abweichungen der objektweisen Anteile fehlender Distanzen, wie dies im Fall zufällig fehlender Daten zu erwarten ist, kann im Vergleich zur Festlegung eines einzigen Werts für die Schrittweite eine deutlichere Verbesserung des Stresswerts der resultierenden Konfiguration erzielt werden.

Neben den bekannteren Zwei-Wege-Verfahren der multidimensionalen Skalierung, d.h. die Datengrundlage ist zweidimensional wie beispielsweise bei einer Distanzmatrix, gibt es auch Drei-Wege-Verfahren. Als dritte Dimension kommen dabei unter anderem die zeitliche Entwicklung der Objektdistanzen oder die Beurteilung der Objektdistanzen durch mehrere Personen in Betracht. Mit den Drei-Wege-Verfahren der multidimensionalen Skalierung werden dann die zeitlichen bzw. individuellen Abweichungen bezüglich der Objektdistanzen herausgearbeitet und bei der Konfiguration der Objekte berücksichtigt (vgl. z.B. *Davison, 1983, S. 121-136*). Damit unterscheiden sich diese Verfahren von einem anderen denkbaren Ansatz, bei dem durch die Mittelung der paarweisen Distanzen über die Zeit bzw. über die beurteilenden Personen letztendlich ein zweidimensionales Datenmaterial resultiert. Im Fall fehlender Distanzen zu einzelnen Zeitpunkten bzw. bei einzelnen Personen ist dann beispielsweise die Anwendung einer Eliminierungsstrategie, d.h. die Mittelwertbildung über die jeweils vorhandenen Distanzen in Betracht zu ziehen.

Ein bei Vorliegen fehlender Distanzen geeignetes Drei-Wege-Verfahren der multidimensionalen Skalierung stellt der von *Takane et al. (1977)* entwickelte Algorithmus **ALSCAL** dar. Dieses Verfahren berücksichtigt die fehlenden Daten auf die gleiche Weise wie die Zwei-Wege-Verfahren der nonmetrischen multidimensionalen Skalierung und basiert ebenfalls nur auf den vorhandenen Distanzen (*MacCallum, 1979, S. 69-70*). Auf eine ausführliche Darstellung dieses Ansatzes wird an dieser Stelle verzichtet. Simulati-

onsstudien zur Leistungsfähigkeit von ALSCAL bei Vorliegen fehlender Daten, deren Ergebnisse grundsätzlich mit den weiter oben bereits dargestellten Resultaten der Simulationsstudien im Fall der Zwei-Wege-Verfahren übereinstimmen, können den Arbeiten von *MacCallum (1979)* und *Malhotra et al. (1988)* entnommen werden.

4.4.3 Faktorenanalyse

Die Methoden der Faktorenanalyse setzen sich das Ziel, die mit den Merkmalen einer Datenmatrix gegebenen Informationen für die Objekte auf möglichst wenige, neue Merkmale, sogenannte Faktoren, zu reduzieren. Im Rahmen der in diesem Abschnitt ausschließlich betrachteten **Hauptkomponentenanalyse** wird eine quantitative Datenmatrix $A=(a_{ik})_{n,m}$ in eine neue Matrix $X=(x_{ik})_{n,r}$ überführt ($r \leq m$), wobei jeder Merkmalsvektor a_k bzw. jede Ausprägung a_{ik} sich als Linearkombination der **Faktoren** $x_1,\ldots,x_r$ bzw. der **Faktorwerte** $x_{i1},\ldots,x_{ir}$ darstellen läßt, d.h.

$$a_k = f_{k1}x_1+\ldots+f_{kr}x_r \quad \text{bzw.} \quad a_{ik} = f_{k1}x_{i1}+\ldots+f_{kr}x_{ir} \quad (i \in N, k \in M). \tag{4.60}$$

Die **Faktorladungen** $f_{k1},\ldots,f_{kr}$ sind dabei die Gewichte, mit denen die Faktoren $x_1,\ldots,x_r$ bzw. Faktorwerte $x_{i1},\ldots,x_{ir}$ in den Merkmalsvektor a_k bzw. in die Ausprägung a_{ik} eingehen. Um bei dieser Überführung einen möglichst geringen Informationsverlust zu gewährleisten, ist das Optimierungsproblem

$$\sum_{i=1}^{n}\sum_{k=1}^{m}\left(a_{ik}-\sum_{l=1}^{r}f_{kl}x_{il}\right)^2 \rightarrow \min \tag{4.61}$$

zu lösen. Nach dem **Ansatz von *Wiberg*** *(1976, S. 230)* ergibt sich im Fall fehlender Daten das folgende, entsprechend modifizierte Optimierungsproblem:

$$\sum_{i,k:\,v_{ik}=1}\left(a_{ik}-\sum_{l=1}^{r}f_{kl}x_{il}\right)^2 \rightarrow \min \tag{4.62}$$

Danach sind die Faktorwerte und Faktorladungen so zu wählen, daß lediglich im Hinblick auf die vorhandenen Daten die Summe der quadrierten Abweichungen zwischen den Ausprägungen und der Linearkombination der entsprechenden Faktorwerte minimal wird. Die Lösung dieses Optimierungsproblems kann jedoch nicht, wie im Fall vollständiger Daten, unter Anwendung der Approximationseigenschaft der Singulärwertzerlegung der Matrix A erfolgen. Statt dessen stellt jeder Algorithmus zur Bestimmung nichtlinearer Kleinst-Quadrate-Schätzungen eine grundsätzlich geeignete Lösungsmethode dar. Aufgrund der Vielzahl der zu schätzenden Parameter schlägt *Wiberg (1976, S. 231)* eine Lösung auf Basis des **Gauß-Newton-Verfahrens** und damit eine Lineari-

sierung des Problems vor. Dabei wird der Gauß-Newton-Algorithmus in zwei Schritten angewandt. Ausgehend von einer gewählten Startlösung für die Faktorladungen werden im ersten Schritt die entsprechenden Faktorwerte der dann linearen Zielfunktion nach dem Kleinst-Quadrate-Prinzip geschätzt. Mit diesen Schätzungen erfolgt im zweiten Schritt eine Kleinst-Quadrate-Schätzung der Faktorladungen, wobei die zu minimierende Zielfunktion aufgrund der Verwendung der im ersten Schritt geschätzten Faktorwerte wiederum eine lineare Form besitzt. Die beiden Schritte werden schließlich bis zur Konvergenz aller zu schätzenden Parameter wiederholt. Mit dem nachfolgenden Beispiel soll nun die Anwendung der Methode von *Wiberg* zur Durchführung einer Hauptkomponentenanalyse auf Basis der vorhandenen Daten verdeutlicht werden.

Beispiel:

Für die Datenmatrix des Anhangs A (Fall 1) werden zur Vereinfachung lediglich die beiden Merkmale Spezialgebiete (7) und Statistische Grafiken (9).betrachtet. Mit dem Ansatz von *Wiberg* soll im folgenden die erste Hauptkomponente bestimmt werden. Damit ergibt sich gemäß (4.62) das Optimierungsproblem

$$\sum_{\substack{i,k:\,v_{ik}=1,\\ i\in N,k\in\{7,9\}}} (a_{ik} - f_k x_i)^2 \;\rightarrow\; \min\,,$$

wobei aufgrund der Anzahl der zu bestimmenden Faktoren von $r = 1$ auf den Index für die Faktorbezeichnung verzichtet wird. Als Startlösung für die Faktorladungen wird $f_7^{(0)} = f_9^{(0)} = \sqrt{0.5}$ gewählt. Durch Nullsetzen der ersten partiellen Ableitungen der Zielfunktion nach den Faktorwerten ergeben sich auf Basis der Faktorladungen $f_7^{(0)}$ und $f_9^{(0)}$ die Faktorwerte $x_1^{(0)},\ldots,x_{15}^{(0)}$ wie folgt, wobei die Berechnung beispielhaft für die Werte $x_1^{(0)}$, $x_4^{(0)}$ und $x_{12}^{(0)}$ dargestellt ist:

$$\left(x_1^{(0)},\ldots,x_{15}^{(0)}\right) = (48.55,7.95,113.14,3.76,41.89,106.12,55.09,55.62,102.71,51.44,60.33,59.40,26.87,30.05,120.56)\,,$$

$$x_1^{(0)} = \frac{\sqrt{0.5}\cdot 48.66 + \sqrt{0.5}\cdot 20.00}{\left(\sqrt{0.5}\right)^2 + \left(\sqrt{0.5}\right)^2} = 48.55\,,\quad x_4^{(0)} = \frac{\sqrt{0.5}\cdot 2.66}{\left(\sqrt{0.5}\right)^2} = 3.76\,,\quad x_{12}^{(0)} = \frac{\sqrt{0.5}\cdot 42.00}{\left(\sqrt{0.5}\right)^2} = 59.40\,.$$

Durch Nullsetzen der ersten partiellen Ableitungen der Zielfunktion nach den Faktorladungen ergeben sich dann auf Basis von $x_1^{(0)},\ldots,x_{15}^{(0)}$ die verbesserten Schätzungen $f_7^{(1)}$ und $f_9^{(1)}$ für die Faktorladungen:

$$f_7^{(1)} = \frac{48.66\cdot 48.55 + 0.00\cdot 7.95 + \ldots + 42.66\cdot 60.33 + 22.00\cdot 26.87}{48.55^2 + 7.95^2 + \ldots + 60.33^2 + 26.87^2} = 0.7895\,,$$

$$f_9^{(1)} = \frac{20.00\cdot 48.55 + \ldots + 83.00\cdot 113.14 + 28.24\cdot 41..89 + \ldots + 20.75\cdot 51.44 + 42.00\cdot 59.40 + \ldots + 85.25\cdot 120.56}{48.55^2 + \ldots + 113.14^2 + 41.89^2 + \ldots + 51.44^2 + 59.40^2 + \ldots + 120.56^2} = 0.6434\,.$$

Diese beiden Schritte sind solange zu wiederholen, bis Konvergenz bezüglich aller zu schätzenden Parameter erreicht wird. In diesem Beispiel resultieren nach dem zehnten Iterationsschritt keine weiteren Veränderungen bei den zu schätzenden Parametern. Als Endergebnis erhält man die folgenden Faktorladungen und Faktorenwerte, wobei das Symbol ^ andeutet, daß es sich um Schätzwerte handelt:

$$\left(\hat{f}_7,\hat{f}_9\right) = (0.8117,0.6239)\,,$$

$$\left(\hat{x}_1,\ldots,\hat{x}_{15}\right) = (49.59,6.70,109.04,3.28,40.82,103.73,56.35,51.78,103.31,52.62,52.56,67.32,26.56,34.06,136.65)\,.$$

Auf die denkbare Normierung des Vektors der Faktorladungen auf die Länge 1 und die Berechnung der daraus resultierenden Faktorwerte wird an dieser Stelle verzichtet. Als Minimum der Zielfunktion ergibt sich ein Wert von 2009.92, d.h. die mittlere quadrierte Abweichung zwischen den vorhandenen Ausprägungen und der Linearkombination der entsprechenden Faktorwerte beträgt $\frac{1}{25} \cdot 2009.92 = 80.40$.

Aufgrund der fehlenden Daten kann der Erklärungsanteil für die erste Hauptkomponente nicht adäquat bestimmt werden. Nimmt man gewisse Verzerrungen in Kauf, dann kann als Erklärungsanteil der Quotient aus der Summe der Merkmalsvarianzen, die auf Basis der vorhandenen Daten berechnet werden, und der Varianz der ersten Hauptkomponente herangezogen werden. Gemäß diesem Ansatz ergibt sich für die erste Hauptkomponente ein Erklärungsanteil von 79.37 Prozent.

Für den von *Wiberg* vorgeschlagenen Ansatz zur Durchführung einer Hauptkomponentenanalyse auf Basis der vorhandenen Daten sind abschließend noch zwei Kritikpunkte festzuhalten. Neben dem aufgrund der fehlenden Daten vorliegenden Mangel, daß die Erklärungsanteile für die einzelnen Faktoren nicht adäquat bestimmt werden können, stellt sich vor allem die Problematik, daß die Bestimmungsgleichungen für die Faktorladungen und Faktorwerte speziell durch das jeweils vorliegende Datenmaterial determiniert sind. Damit können die Bestimmungsgleichungen nicht in einer einfachen, matriziellen Form angegeben werden.

4.4.4 Regressionsanalyse

Einen im Fall fehlender Daten modifizierten Ansatz einer einfachen Regression stellt die **Methode von *Afifi und Elashoff*** *(1967, S. 15-16)* dar. Ausgehend von einem abhängigen Merkmal $k \in M$ und einem unabhängigen Merkmal $l \in M$, $l \neq k$ mit jeweils fehlenden Daten sind die Regressionskoeffizienten β_0 und β_1 der Regressionsfunktion

$$a_{ik} = \beta_0 + \beta_1 a_{il} \quad (i \in N, k,l \in M, l \neq k) \tag{4.63}$$

zu bestimmen. Dies erfolgt durch die Minimierung einer im Hinblick auf die fehlenden Daten modifizierten Zielfunktion nach dem Kleinst-Quadrate-Prinzip. Dabei sind die Regressionskoeffizienten β_0 und β_1 sowie die für alle fehlenden Ausprägungen der beiden Merkmale jeweils heranzuziehenden Imputationswerte $a_{\bullet k}$ und $a_{\bullet l}$ so zu wählen, daß die Zielfunktion

$$\sum_{i \in N_{kl}} (a_{ik} - \beta_0 - \beta_1 a_{il})^2 + \sum_{i \in \overline{N}_k \cap N_l} (a_{\bullet k} - \beta_0 - \beta_1 a_{il})^2 + \sum_{i \in N_k \cap \overline{N}_l} (a_{ik} - \beta_0 - \beta_1 a_{\bullet l})^2 \tag{4.64}$$

minimal wird. Dabei ist $N_{kl} = \{i: v_{ik} = 1 \wedge v_{il} = 1\}$, $N_k = \{i: v_{ik} = 1\}$ und $\overline{N}_k$ die Komplementärmenge von N_k bezüglich N. Durch Nullsetzen der ersten partiellen Ableitungen von (4.64) nach β_0, β_1, $a_{\bullet k}$ und $a_{\bullet l}$ ergeben sich dann die folgenden Schätzungen für die

Regressionskoeffizienten und die Imputationswerte, wobei das Symbol ^ andeutet, daß es sich jeweils um Schätzwerte handelt:

$$\hat{\beta}_1 = \frac{\sum_{i \in N_{kl}} \left(a_{ik} - \frac{1}{|N_{kl}|} \sum_{i \in N_{kl}} a_{ik} \right) \left(a_{il} - \frac{1}{|N_{kl}|} \sum_{i \in N_{kl}} a_{il} \right)}{\sum_{i \in N_{kl}} \left(a_{il} - \frac{1}{|N_{kl}|} \sum_{i \in N_{kl}} a_{il} \right)^2 + \sum_{i \in \overline{N}_k \cap N_l} \left(a_{il} - \frac{1}{|\overline{N}_k \cap N_l|} \sum_{i \in \overline{N}_k \cap N_l} a_{il} \right)^2}, \tag{4.65}$$

$$\hat{\beta}_0 = \frac{1}{|N_{kl}|} \sum_{i \in N_{kl}} a_{ik} - \hat{\beta}_1 \frac{1}{|N_{kl}|} \sum_{i \in N_{kl}} a_{il}, \tag{4.66}$$

$$\hat{a}_{\bullet k} = \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{|\overline{N}_k \cap N_l|} \sum_{i \in \overline{N}_k \cap N_l} a_{il}, \quad \hat{a}_{\bullet l} = \frac{-\hat{\beta}_0 + \frac{1}{|N_k \cap \overline{N}_l|} \sum_{i \in N_k \cap \overline{N}_l} a_{ik}}{\hat{\beta}_1}. \tag{4.67}$$

Die Schätzungen der Imputationswerte für die fehlenden Ausprägungen der Merkmale k und l gemäß (4.67), die nach diesem Ansatz heranzuziehen sind, werden lediglich zur Bestimmung der Schätzwerte für die Regressionskoeffizienten benötigt und mit den nach (4.65) und (4.66) angegebenen Gleichungen berücksichtigt. Eine explizite Berechnung dieser Imputationswerte ist somit nicht erforderlich.

Beispiel:

Für die Datenmatrix des Anhangs A (Fall 1) werden im folgenden lediglich die beiden Merkmale Spezialgebiete (7) und Statistische Grafiken (9) betrachtet. Wird das Merkmal Spezialgebiete als abhängige und das Merkmal Statistische Grafiken als unabhängige Variable herangezogen, dann sind die Regressionskoeffizienten β_0 und β_1 der Regressionsfunktion

$$a_{i7} = \beta_0 + \beta_1 a_{i9} \quad (i \in N)$$

zu schätzen. Mit $N_{79} = \{1,2,3,5,6,7,8,9,10,13\}$, $\overline{N}_7 = \{12,14,15\}$ und $N_9 = \{1,2,3,5,6,7,8,9,10,12,13,14,15\}$ ergeben sich nach dem Ansatz von *Afifi und Elashoff* unter Verwendung von (4.65) und (4.66) die folgenden Schätzwerte für die Regressionskoeffizienten:

$$\hat{\beta}_1 = \frac{(48.66-48.83)(20.00-37.35)+\ldots+(22.00-48.83)(16.00-37.35)}{(20.00-37.35)^2+\ldots+(16.00-37.35)^2+(42.00-49.50)^2+(21.25-49.50)^2+(85.25-49.50)^2} = 0.587,$$

$$\hat{\beta}_0 = 48.83 - 0.587 \cdot 37.35 = 26.91.$$

Auf die explizite Berechnung der Imputationswerte nach (4.67), die für die fehlenden Ausprägungen der beiden Merkmale jeweils heranzuziehen sind, wird an dieser Stelle verzichtet.

Der dargestellte Ansatz ist grundsätzlich auch auf Regressionsmodelle mit mehr als einer unabhängigen Variable übertragbar. Mit zunehmender Anzahl der Regressoren werden jedoch die analog zu (4.64) zu modifizierende Zielfunktion und damit auch die

resultierenden Gleichungen zur Bestimmung der Schätzwerte für die entsprechenden Regressionskoeffizienten immer komplexer. Auf eine ausführliche Darstellung wird an dieser Stelle verzichtet.

4.4.5 Vergleich der Verfahren

Alle in den Abschnitten 4.4.1 bis 4.4.4 betrachteten MD-Verfahren beruhen auf einer Modifikation multivariater Analyseverfahren zur Berücksichtigung fehlender Daten, wobei die in der Datenanalyse relevanten Aufgabenstellungen der Klassifikation, Repräsentation und Identifikation behandelt wurden. In der Tabelle 4.5 sind die Voraussetzungen sowie die wichtigsten Eigenschaften der vorgestellten Verfahren noch einmal zusammenfassend dargestellt.

Verfahren	**Voraussetzungen**	**Eigenschaften**
MVL Verfahren	Distanzmatrix, Daten sind MCAR	Bestimmung einer hierarchischen Klassifikation auf Basis der vorhandenen Distanzen durch Modifikation des jeweiligen Verschiedenheitsindexes
PACII Verfahren	Distanzmatrix, Daten sind MCAR	Bestimmung einer pyramidalen Klassifikation auf Basis der vorhandenen Distanzen durch Modifikation des jeweiligen Verschiedenheitsindexes
Verfahren von Kruskal auf Basis der vorhandenen Distanzen	Distanzmatrix, Daten sind MCAR	Bestimmung einer adäquaten Anordnung der Objekte in einem möglichst niedrig dimensionierten Raum unter ausschließlicher Berücksichtigung der vorhandenen Distanzen
ALSCAL Verfahren	Dreidimensionale Distanzmatrix, Daten sind MCAR	Drei-Wege-Verfahren der multidimensionalen Skalierung unter ausschließlicher Berücksichtigung der vorhandenen Distanzen
Methode von *Wiberg*	Datenmatrix (quantitativ), Daten sind MCAR	Bestimmung der Hauptkomponenten auf Basis der vorhandenen Daten durch Modifikation der nach dem Kleinst-Quadrate-Prinzip sich ergebenden Zielfunktion
Methode von *Afifi und Elashoff*	Datenmatrix (ein kardinales Merkmal mit MD, ein weiteres kardinales oder dichotomes Merkmal mit MD), Daten sind MAR	Bestimmung der Regressionskoeffizienten auf Basis der vorhandenen Daten durch Modifikation der nach dem Kleinst-Quadrate-Prinzip sich ergebenden Zielfunktion

Tabelle 4.5: Vergleich der modifizierten multivariaten Verfahren

Der gemeinsame Nenner aller betrachteten Verfahren besteht darin, daß auf Basis der vorhandenen Daten die entsprechenden Analyseergebnisse unmittelbar, d.h. ohne die vorherige Anwendung eines Eliminierungs-, Imputations- oder Parameterschätzverfahrens, resultieren. Falls in einer datenanalytischen Untersuchung mehrere Aufgabenstellungen vorliegen und somit unterschiedliche multivariate Verfahren anzuwenden sind, kann das Heranziehen der hier vorgestellten multivariaten Analyseverfahren mit einem sehr hohen Aufwand verbunden sein. Eine Eliminierungs- oder Imputationsstrategie stellt in diesem Fall einen unter Umständen geeigneteren Lösungsansatz zur Berücksichtigung der fehlenden Daten dar.

4.5 Sensitivitätsbetrachtungen

Die bislang vorgestellten Verfahren zur Behandlung fehlender Daten führen jeweils zu einem einzigen Resultat. Dieses liegt abhängig von der gewählten Strategie in Form einer vollständigen Daten- bzw. Distanzmatrix, einer Schätzung bestimmter Parameter oder in Form multivariater Analyseergebnisse vor. In diesem Abschnitt sollen nun Ansätze betrachtet werden, bei denen die Behandlung fehlender Daten zu einer Reihe, unter Umständen auch unterschiedlicher Ergebnisse führt, d.h. es resultieren letztendlich entweder mehrere vollständige Daten- bzw. Distanzmatrizen, mehrere Schätzungen von Parametern oder mehrere Analyseergebnisse. Auf Basis dieser Ergebnisse kann schließlich eine Sensitivitätsbetrachtung in der Art durchgeführt werden, daß die Sensitivität der Ergebnisse gegenüber dem zugrundegelegten Ausfallmodell und dem verwendeten MD-Verfahren beurteilt wird. Dazu wird in **Abschnitt 4.5.1** zunächst die sogenannte multiple Imputation dargestellt, bei der für die fehlenden Daten jeweils mehrere Imputationswerte bestimmt werden. Der **Abschnitt 4.5.2** widmet sich dann weiteren Möglichkeiten einer Sensitivitätsbetrachtung, die im Vergleich zur multiplen Imputation in der Literatur jedoch eine geringere Bedeutung besitzen. In **Abschnitt 4.5.3** erfolgt abschließend eine vergleichende Darstellung der vorgestellten Ansätze. An dieser Stelle ist noch festzuhalten, daß die in der Literatur ausnahmslos von einer unvollständigen Datenmatrix ausgehenden Ansätze einer Sensitivitätsbetrachtung auch auf den Fall einer unvollständigen, unmittelbar erhobenen Distanzmatrix übertragbar sind.

4.5.1 Multiple Imputation

Das Konzept der multiplen Imputation wurde von *Rubin* in einer Reihe von Arbeiten entwickelt und theoretisch begründet (vgl. *Rubin, 1977, 1978, 1979, 1987*). Den Ausgangspunkt dieses Ansatzes stellt die Überlegung dar, daß im Prinzip jede Strategie und damit auch jedes Verfahren zur Behandlung fehlender Daten ein Modell des Ausfallme-

chanismus heranzieht, wobei die meisten MD-Verfahren das Modell eines unsystematischen Ausfallmechanismus verwenden. Auf Basis eines oder auch mehrerer Ausfallmodelle werden im Rahmen einer multiplen Imputation für alle fehlenden Daten jeweils mehrere Imputationswerte bestimmt, so daß anschließend eine Beurteilung der Sensitivität von Analyseergebnissen möglich ist. Eine multiple Imputation besteht grundsätzlich aus den folgenden drei Schritten:

1. Der den Daten zugrundeliegende Ausfallmechanismus ist zu modellieren, wobei ein einziges oder auch mehrere unterschiedliche Modelle des Ausfallmechanismus herangezogen werden können.
2. Für jedes betrachtete Ausfallmodell sind unter Verwendung eines nichtdeterministischen Imputationsverfahrens, wie beispielsweise eines Hot-Deck-Verfahrens oder einer Imputation mittels Zufallsauswahl, jeweils mehrere Imputationswerte für alle fehlenden Daten zu bestimmen. Entsprechend resultieren für jedes Ausfallmodell mehrere vollständige Daten- bzw. Distanzmatrizen.
3. Auf Basis aller im zweiten Schritt bestimmten vollständigen Daten- bzw. Distanzmatrizen wird jeweils eine Datenauswertung gemäß dem Untersuchungsziel durchgeführt. Falls ein einziges Modell des Ausfallmechanismus verwendet wurde, ist eine Simulation der Verteilung einzelner Analyseparameter und damit auch die Bestimmung von entsprechenden Schätzintervallen möglich. Falls mehrere Modelle des Ausfallmechanismus herangezogen wurden, kann durch einen Vergleich der Analyseergebnisse die Sensitivität dieser Ergebnisse gegenüber unterschiedlichen Annahmen über den Ausfallmechanismus beurteilt werden.

Die multiple Imputation liefert also „... keine eindeutigen Resultate, sondern zeigt die Spannweite möglicher Ergebnisse auf." *[Schnell, 1986, S. 224]*

Beispiel:

Für die Datenmatrix des Anhangs A wird zunächst der Fall eines zufälligen Fehlens der Daten (Fall 1) betrachtet. Es wird angenommen, daß dieser den Daten zugrundeliegende Ausfallmechanismus bekannt ist. Gemäß dem Ansatz der multiplen Imputation sind also auf Basis eines Ausfallmechanismus für die fehlenden Ausprägungen jeweils mehrere Imputationswerte zu bestimmen, wobei in diesem Beispiel eine Imputation durch Zufallsauswahl erfolgen soll. Zur Vereinfachung wird im folgenden lediglich das Merkmal Preisniveau (1) betrachtet. In Anlehnung an das erste Beispiel des Abschnitts 4.2.1.3 wird für die Ausprägungen a_{i1} die Wahrscheinlichkeitsfunktion

$$f(a_{i1}) = \begin{cases} \frac{2}{13} & \text{für } a_{i1} = \text{niedrig} \\ \frac{5}{13} & \text{für } a_{i1} = \text{mittel} \\ \frac{2}{13} & \text{für } a_{i1} = \text{gehoben} \\ \frac{4}{13} & \text{für } a_{i1} = \text{hoch} \\ 0 & \text{sonst} \end{cases}$$

unterstellt. Bei zehn Wiederholungen des Imputationsprozesses unter Verwendung eines Pseudozufallszahlengenerators erhält man die in der nachfolgenden Tabelle angegebenen Schätzwerte für die fehlenden Ausprägungen. Ausgehend von den jeweils resultierenden vollständigen Merkmalsvektoren kann als Analyseergebnis beispielsweise der Median bestimmt werden. Die auf Basis der zehn Imputationsläufe sich ergebenden Mediane des Merkmals Preisniveau sind ebenfalls in der folgenden Tabelle dargestellt.

Imputationslauf	1	2	3	4	5	6	7	8	9	10
Schätzwert für a_{21}	mittel	mittel	mittel	gehoben	hoch	mittel	hoch	gehoben	hoch	mittel
Schätzwert für $a_{10,1}$	hoch	gehoben	mittel	hoch	mittel	gehoben	mittel	niedrig	gehoben	mittel
Median	mittel	mittel	mittel	gehoben	mittel	mittel	mittel	mittel	gehoben	mittel

Bei Betrachtung aller Imputationsläufe resultiert als Median die Ausprägung „mittel" in 80 Prozent und die Ausprägung „gehoben" in 20 Prozent aller Fälle. Auf Basis der gegebenen Verteilung der Merkmalsausprägungen können in diesem Beispiel die theoretisch exakten und für eine hinreichend große Anzahl von Imputationsläufen zu erwartenden Werte der Verteilung des Medians relativ einfach bestimmt werden. Diese ergeben sich mit 78.7 Prozent (Ausprägung „mittel") und 21.3 Prozent (Ausprägung „gehoben").

Für die Datenmatrix des Anhangs A wird nun der dort beschriebenen Fall 4 betrachtet, gemäß dem das Fehlen der Ausprägungen beim Merkmal Multivariate Verfahren (6) von den jeweiligen Werten dieser Ausprägungen abhängt. Es wird angenommen, daß zwar die grundsätzlich vorliegende Abhängigkeitsbeziehung der fehlenden Daten bekannt ist, jedoch keine genauen Informationen über den Ausfallmechanismus vorliegen. In diesem Fall sind gemäß dem Konzept der multiplen Imputation mehrere Modelle des Ausfallmechanismus zu berücksichtigen und für jedes dieser Modelle sind für jede fehlende Ausprägung mehrere Imputationswerte zu bestimmen. Zur Illustration werden die folgenden fünf Ausfallmodelle betrachtet, die sich in Anlehnung an das erste Beispiel des Abschnitts 4.2.4 ergeben:

$$P\left(v_{i6}=0 \mid A\right)=\begin{cases}0.2 & \text{für } 0 \le a_{i6} \le c \\ 0.9 & \text{für } c < a_{i6} \le 100 \\ 0 & \text{sonst}\end{cases} \quad \text{mit } c = 40,45,50,55,60\,.$$

Für die Ausprägungen a_{i6} wird im Intervall $[0,100]$ eine Normalverteilung mit $\mu = 35.00$ und $\sigma = 20.00$ angenommen. Bei jeweils zwei Imputationsläufen für jedes Ausfallmodell ergeben sich unter Verwendung eines Pseudozufallszahlengenerators die folgenden Schätzwerte für die fehlenden Ausprägungen:

Ausfallmodell	1 (c = 40)		2 (c = 45)		3 (c = 50)		4 (c = 55)		5 (c = 60)	
Imputationslauf	1	2	1	2	1	2	1	2	1	2
Schätzwert für a_{16}	45.02	48.83	56.00	50.57	53.80	53.78	92.30	61.17	60.15	75.19
Schätzwert für a_{36}	52.53	74.72	32.93	58.23	14.58	50.17	72.77	56.70	15.46	60.36
Schätzwert für a_{96}	62.49	7.17	46.91	53.47	52.70	55.41	68.91	55.71	66.80	62.23
Schätzwert für $a_{10,6}$	46.98	46.97	46.10	46.24	92.94	28.60	64.91	67.42	65.56	85.15
Schätzwert für $a_{15,6}$	75.90	41.68	53.35	50.34	51.38	52.78	23.73	59.13	68.17	46.90

Ausgehend von den jeweils resultierenden vollständigen Merkmalsvektoren kann als Analyseergebnis beispielsweise das arithmetische Mittel des Merkmals Multivariate Verfahren bestimmt werden. Für die durchgeführten Imputationen der fünf Ausfallmodelle ergeben sich dabei die folgenden Werte:

Ausfallmodell	1 ($c = 40$)		2 ($c = 45$)		3 ($c = 50$)		4 ($c = 55$)		5 ($c = 60$)	
Imputationslauf	1	2	1	2	1	2	1	2	1	2
Arithmetisches Mittel	33.71	29.47	30.53	32.10	32.54	30.89	36.35	34.85	33.25	36.83

Aufgrund der geringen Anzahl an durchgeführten Imputationsläufen für die einzelnen Ausfallmodelle sind die Ergebnisse nicht sehr aussagefähig. Das zu erwartende Ergebnis, daß das arithmetische Mittel zunimmt, wenn der in den Ausfallmodellen variierte Parameter c größer wird, kann lediglich teilweise bestätigt werden. Dennoch wird die Spannweite der möglichen Ergebnisse und damit die Sensitivität des arithmetischen Mittels gegenüber dem Ausfallmodell ersichtlich.

Das Konzept der multiplen Imputation weist eine Reihe von Nachteilen auf. So ist eine praktische Anwendung im allgemeinen sehr aufwendig, da eine Vielzahl vollständiger Daten- bzw. Distanzmatrizen zu bestimmen und die entsprechenden Analyseverfahren auf Basis dieser Matrizen anzuwenden sind (vgl. *Herzog, Rubin, 1983, S. 242*). Des weiteren ist eine Modellierung des Ausfallmechanismus sowie ein Vergleich der resultierenden Analyseergebnisse mit Problemen behaftet. Jedoch besitzt die multiple Imputation gegenüber allen MD-Verfahren den entscheidenden Vorteil, daß ohne die genaue Kenntnis des zugrundeliegenden Ausfallmechanismus das Problem fehlender Daten in seinen möglichen Auswirkungen abgeschätzt werden kann (*Schnell, 1986, S. 225*).

4.5.2 Weitere Ansätze

In diesem Abschnitt sollen noch kurz drei weitere Ansätze einer Sensitivitätsbetrachtung im Fall fehlender Daten vorgestellt werden, die im Vergleich zur multiplen Imputation jedoch eine geringere Bedeutung besitzen.

Einen denkbaren Ansatz einer Sensitivitätsbetrachtung stellt die **multiple Anwendung von MD-Verfahren** dar. Unter Berücksichtigung der jeweiligen Voraussetzungen kann durch die Anwendung mehrerer unterschiedlicher MD-Verfahren die Sensitivität von Imputationswerten, geschätzten Parametern oder datenanalytischen Ergebnissen untersucht werden. So können beispielsweise einfache und multivariate Imputationstechniken herangezogen werden, um dadurch die Spannweite der denkbaren Imputationswerte aufzuzeigen. Die Verwendung unterschiedlicher Ersetzungstechniken, wie *Schnell (1986, S. 227-228)* bemerkt, stellt jedoch keine multiple Imputation dar. Des weiteren kann die Sensitivität eines Analyseergebnisses durch die Anwendung unterschiedlicher Strategien beurteilt werden. Als Beispiel sei eine Faktorenanalyse bei Vor-

liegen einer unvollständigen Datenmatrix genannt, deren Ergebnisse auf Basis eines Eliminierungsverfahrens, eines Imputationsverfahrens sowie des entsprechenden multivariaten Analyseverfahrens bestimmt und anschließend miteinander verglichen werden können.

Ein weiterer Ansatz besteht in der Verwendung sogenannter **Extremmodelle**, d.h. auf Basis der vorliegenden Daten werden extreme Ausfallmodelle entwickelt, so daß die Ergebnisse, die durch Anwendung eines geeigneten MD-Verfahrens resultieren, entsprechende Extremwerte darstellen (vgl. z.B. *Schnell, 1986, S. 226*). Werden als Imputationswerte für die fehlenden Daten beispielsweise die minimal und maximal denkbaren Werte herangezogen, dann können die Analyseergebnisse, die sich auf Basis der daraus resultierenden Daten- bzw. Distanzmatrizen ergeben, anschließend miteinander verglichen werden. Da bei diesem Ansatz keine realistische Modellierung des Ausfallmechanismus erfolgt, handelt es sich auch im Fall einer Verwendung von Imputationsverfahren um keine multiple Imputation.

Ein abschließend zu erwähnender Ansatz einer Sensitivitätsbetrachtung ist das von ***Simon und Simonoff*** *(1986)* im Rahmen einer multiplen Regression vorgeschlagene Verfahren, mit dem eine Abschätzung für die Regressionskoeffizienten auf Basis eines unbekannten Ausfallmechanismus erfolgt. Als Resultat erhält man obere und untere Grenzen für die einzelnen Regressionskoeffizienten in Abhängigkeit einer Maßzahl für die Art des Ausfallmechanismus. Damit ergibt sich letztendlich ein Bereich von möglichen Werten für die Regressionskoeffizienten über alle denkbaren Ausfallmechanismen hinweg. Dieser grundsätzlich sehr interessante Ansatz ist jedoch im Hinblick auf reale MD-Probleme aus den folgenden Gründen ungeeignet: Zum einen beschränkt sich das Verfahren auf den Fall, daß lediglich eine einzige unabhängige Variable fehlende Daten aufweist. Zum anderen müssen unendlich viele, theoretisch denkbare Ausfallmechanismen in Form ebenfalls unendlich vieler, entsprechender Maßzahlen berücksichtigt werden. Somit resultieren im allgemeinen sehr große Bereiche für möglichen Werte der Regressionskoeffizienten. Dies wird auch anhand der bei *Simon und Simonoff* dargestellten Beispiele deutlich.

4.5.3 Vergleich der Verfahren

Alle vorgestellten Ansätze einer Sensitivitätsbetrachtung verfolgen grundsätzlich das Ziel, die Spannweite der im Rahmen der Behandlung fehlender Daten resultierenden Ergebnisse aufzuzeigen. Abhängig vom gewählten Ansatz erhält man mehrere vollständige Daten- bzw. Distanzmatrizen, mehrere Parameterschätzungen oder mehrere Analyseergebnisse, so daß eine Beurteilung der Sensitivität dieser Ergebnisse gegenüber

dem zugrundegelegten Ausfallmodell und dem verwendeten MD-Verfahren möglich ist. In der Tabelle 4.6 sind die Voraussetzungen sowie die wichtigsten Eigenschaften der in den Abschnitten 4.5.1 und 4.5.2 behandelten Ansätze einer Sensitivitätsbetrachtung noch einmal zusammengefaßt.

Verfahren	**Voraussetzungen**	**Eigenschaften**
Multiple Imputation	Daten- oder Distanzmatrix, Modell des Ausfallmechanismus	Bestimmung mehrerer Imputationswerte für die fehlenden Daten, Beurteilung der Sensitivität von Analyseergebnissen gegenüber dem Ausfallmodell und dem MD-Verfahren
Multiple Anwendung von MD-Verfahren	Daten- oder Distanzmatrix, weitere Voraussetzung gemäß der verwendeten MD-Verfahren	Verwendung mehrerer MD-Verfahren, Beurteilung der Sensitivität von Imputationswerten, Parameterschätzungen oder Analyseergebnissen gegenüber den betrachteten MD-Verfahren
Sensitivitätsbetrachtung mit Extremmodellen	Daten- oder Distanzmatrix	Verwendung von extremen Ausfallmodellen, Beurteilung der Sensitivität von Imputationswerten, Parameterschätzungen oder Analyseergebnissen gegenüber den betrachteten Extremmodellen
Methode von *Simon und Simonoff*	Datenmatrix	Bestimmung der Spannweite für die Regressionskoeffizienten im Rahmen einer multiplen Regression in Abhängigkeit des unbekannten Ausfallmechanismus

Tabelle 4.6: Vergleich der Ansätze im Rahmen einer Sensitivitätsbetrachtung

Alle in der Tabelle 4.6 dargestellten Ansätze besitzen gegenüber der Verwendung eines einzigen MD-Verfahrens zwar den Vorteil, daß die Auswirkungen der zugrundegelegten Ausfallmechanismen bzw. der verwendeten MD-Verfahren näher analysiert und abgeschätzt werden können. Dem gegenüber stellt jedoch der im allgemeinen sehr hohe Aufwand, der mit einer praktischen Anwendung dieser Ansätze verbunden ist, einen entscheidenden Nachteil dar. Damit läßt sich auch die relativ geringe Bedeutung erklären, die den Ansätzen einer Sensitivitätsbetrachtung in der Literatur beigemessen wird, obwohl diese Ansätze im Fall eines unbekannten Ausfallmechanismus - ein Problem, das gerade in realen datenanalytischen Untersuchungen häufig auftritt - die einzige Möglichkeit zur Abschätzung der Konsequenzen für die Auswertung einer unvollständigen Daten- oder Distanzmatrix darstellen.

5 Zusammenfassung und Ausblick

Den Ausgangspunkt dieser Arbeit stellte eine unvollständige Datenmatrix der Form (1.1) oder eine unvollständige, unmittelbar erhobene Distanzmatrix der Form (1.3) dar. Wie in Kapitel 2 zunächst aufgezeigt wurde, können fehlende Daten auf eine Reihe unterschiedlicher Ursachen zurückzuführen sein. Dabei sind lediglich die Auswirkungen der möglichen Ausfallursachen auf das vorliegende Datenmaterial von Bedeutung. Diese Tatsache wird durch die Definition systematischer bzw. unsystematischer Ausfallmechanismen berücksichtigt. Da der den Daten zugrundeliegende Ausfallmechanismus Konsequenzen für die Auswertung der Daten hat, werden zur adäquaten Behandlung der fehlenden Werte Kenntnisse über den vorliegenden Ausfallmechanismus benötigt. Damit ergibt sich die Notwendigkeit einer Strukturanalyse der unvollständigen Daten- bzw. Distanzmatrix, die in Kapitel 3 dieser Arbeit ausführlich dargestellt wurde. Die Ergebnisse einer derartigen Analyse dienen dazu, entweder die Annahme eines unsystematischen Ausfallmechanismus zu rechtfertigen oder aber einen systematischen Ausfallmechanismus transparenter zu machen. Entsprechend dem zugrundeliegenden Ausfallmechanismus können zur Behandlung der unvollständigen Daten- bzw. Distanzmatrix schließlich die in Kapitel 4 dieser Arbeit vorgestellten Methoden, die sich gemäß der grundsätzlich verfolgten Strategie in die fünf Verfahrenskategorien Eliminierungsverfahren, Imputationsverfahren, Parameterschätzverfahren, multivariate Analyseverfahren sowie Sensitivitätsbetrachtungen einteilen lassen, zur Anwendung kommen.

Die grundlegende Vorgehensweise im Fall fehlender Daten besteht somit zunächst aus einer Strukturanalyse der unvollständigen Daten- bzw. Distanzmatrix sowie der anschließenden Anwendung eines MD-Verfahrens, das gemäß dem Ergebnis der Strukturanalyse in Betracht zu ziehen ist. Dabei stellen sich jedoch noch die folgenden Fragen:

- Welche Strategie bzw. welches MD-Verfahren ist unter bestimmten Voraussetzungen geeigneter als andere Strategien bzw. MD-Verfahren ?
- Kann ein auf einem unsystematischen Ausfallmechanismus der Daten basierendes MD-Verfahren auch bei Vorliegen eines systematischen Ausfallmechanismus angewandt werden ?
- Welche Möglichkeiten der EDV-gestützten Auswertung einer unvollständigen Daten- bzw. Distanzmatrix gibt es ?

In der Literatur existieren eine Vielzahl von empirischen Untersuchungen und Simulationsstudien, die der Frage nach der Effizienz der einzelnen Strategien bzw. MD-Verfahren nachgehen. Empirische Untersuchungen gehen dabei von einem realen unvollständigen Datenmaterial aus und haben somit den Nachteil, daß die fehlenden Daten unbe-

kannt sind. In Simulationsstudien hingegen werden die als fehlend zu betrachtenden Daten aus dem vollständigen Datenmaterial ausgewählt, so daß anschließend, je nach Untersuchungsziel, ein Vergleich mit den tatsächlichen Daten oder den auf den vollständigen Daten basierenden Analyseergebnissen möglich ist. Dabei können unterschiedliche Gegebenheiten des Datenmaterials simuliert und somit unterschiedliche Einflußfaktoren, wie beispielsweise der Anteil fehlender Daten, die Interkorrelationen der Merkmale oder der zugrundegelegte Ausfallmechanismus, bezüglich der Effizienz der MD-Verfahren untersucht werden.

Einige Simulationsstudien im Fall einer unvollständigen Datenmatrix einschließlich des jeweiligen Untersuchungsdesigns sowie der Ergebnisse können der Tabelle 5.1 entnommen werden. Dabei ist grundsätzlich festzuhalten, daß ein durchgängig, d.h. bei unterschiedlichen Gegebenheiten einer unvollständigen Datenmatrix jeweils bestes MD-Verfahren nicht existiert. Lediglich unter gewissen Bedingungen kann ein Verfahren bessere Ergebnisse als andere Verfahren liefern. So haben beispielsweise *Kim und Curry (1977)* in ihrer Studie gezeigt, daß bei Vorliegen von MCAR sowie geringer Interkorrelationen der Merkmale die Auswertung der jeweils verfügbaren Objekte durchgängig bessere Ergebnisse liefert als eine Auswertung der vollständig erhobenen Objekte. Das Vorliegen hoher Interkorrelationen der Merkmale führt jedoch zur Umkehrung dieses Resultats (vgl. *Haitovsky, 1968, S. 79*). Dabei ist jedoch zu beachten, daß in einer Simulationsstudie die Daten so erzeugt werden, daß die jeweils gewünschten Interkorrelationen vorliegen. In realen Untersuchungen ist es lediglich möglich, die Korrelationen der Merkmale auf Basis der vorhandenen Daten empirisch zu ermitteln. Folglich muß angenommen werden, daß die fehlenden Daten den gleichen Gesetzmäßigkeiten unterliegen, d.h. die anhand der vorhandenen Daten festgestellten und die im Fall vollständiger Daten vorliegenden Korrelationen gleich sind.

Grundsätzlich entsprechen die Ergebnisse der einzelnen Studien den Erwartungen. Im Fall mittlerer und hoher Interkorrelationen liefern die auf den Zusammenhängen zwischen den Merkmalen basierenden MD-Verfahren die besten Ergebnisse. Sind die Korrelationen lediglich niedrig, ist die Mittelwertimputation den multivariaten Imputationstechniken überlegen. Die Eliminierungsverfahren führen meist zu den schlechtesten Resultaten, da sie keinerlei im Datenmaterial enthaltene Information nutzen und eine Anwendung ohnehin nur bei einer geringen Anzahl fehlender Daten geeignet ist. Mit zunehmendem Anteil fehlender Daten sinkt auch die Effizienz aller jeweils untersuchten MD-Verfahren. Eine Robustheit gegenüber dem Ausfallmechanismus kann für keines der jeweils untersuchten MD-Verfahren festgestellt werden. Damit führt die Anwendung eines MD-Verfahrens, das auf einem unsystematischen Ausfallmechanismus basiert, bei Vorliegen eines systematischen Ausfallmechanismus zu einer unter Umständen erheblichen Verzerrung der Ergebnisse.

Untersuchung von	Untersuchungsdesign	Ergebnisse
Haitovsky (1968)	**MD-Verfahren:** Complete-case und available-case analysis **Einflußfaktoren:** Stichprobenumfang, MD-Anteil, Interkorrelation der Merkmale, Ausfallmechanismus **Untersuchungsziel:** Berechnung von Regressionskoeffizienten	Die complete-case analysis ist der available-case analysis, außer im Fall hoher Interkorrelationen, überlegen. Beide Verfahren sind bei hohem MD-Anteil oder systematischem Ausfallmechanismus ungeeignet.
Timm (1970)	**MD-Verfahren:** Complete-case analysis, Mittelwertimputation, Methoden nach *Dear* und *Buck* **Einflußfaktoren:** Stichprobenumfang, MD-Anteil, Interkorrelation der Merkmale, Anzahl der Merkmale **Untersuchungsziel:** Berechnung einer Korrelations- bzw. einer Kovarianzmatrix	Kein MD-Verfahren liefert unter allen Bedingungen die besten Ergebnisse. Bei mittleren und hohen Interkorrelationen ist abhängig vom MD-Anteil die Methode nach *Dear* oder *Buck* vorzuziehen. Die Mittelwertimputation ist bei geringen Interkorrelationen geeignet.
Chan und Dunn (1974)	**MD-Verfahren:** Complete-case und available-case analysis, Mittelwertimputation, Methoden nach *Dear*, *Buck* und *Walsh* **Einflußfaktoren:** Stichprobenumfang, MD-Anteil, Interkorrelation der Merkmale, Anzahl der Merkmale **Untersuchungsziel:** Durchführung einer Diskriminanzanalyse	Bei mittleren bis hohen Interkorrelationen führen die Methoden nach *Dear* und *Buck*, bei niedrigen Interkorrelationen die Mittelwertimputation zu den besten Ergebnissen. Eine Vergrößerung des Stichprobenumfangs oder eine Reduzierung des MD-Anteils führt allgemein zu einer Ergebnisverbesserung.
Kim und Curry (1977)	**MD-Verfahren:** Complete-case und available-case analysis **Einflußfaktoren:** MD-Anteil, Anzahl der Merkmale **Untersuchungsziel:** Berechnung einer Korrelationsmatrix	Für die in der Untersuchung nicht variierte, niedrig gewählte Interkorrelation der Merkmale liefert die available-case analysis durchgängig bessere Ergebnisse als die complete-case analysis.
Schnell (1985)	**MD-Verfahren:** Complete-case und available-case analysis, Mittelwertimputation, Hot-Deck-Verfahren, EM-Algorithmus, Methoden nach *Dear* und *Buck* **Einflußfaktoren:** Stichprobenumfang, MD-Anteil, Interkorrelation der Merkmale, Ausfallmechanismus **Untersuchungsziel:** Berechnung einer Korrelationsmatrix bzw. Bestimmung von Imputationswerten	Kein MD-Verfahren liefert unter allen Bedingungen die besten Ergebnisse. Vor allem bei mittleren bis hohen Interkorrelationen liefern die Methoden nach *Dear* und *Buck* sowie der EM-Algorithmus im allgemeinen die besseren Resultate. Eine Robustheit gegenüber dem Ausfallmechanismus ist jedoch bei keinem der untersuchten MD-Verfahren feststellbar.

Tabelle 5.1: Überblick einiger Simulationsstudien zur Effizienz von MD-Verfahren

Die Entscheidung zugunsten einer Strategie bzw. eines MD-Verfahrens hängt aber nicht nur von den an das Datenmaterial gestellten Anforderungen ab, sondern wird auch von der Zielsetzung der Analyse determiniert. So kann beispielsweise die Anwendung eines multivariaten Analyseverfahrens durchaus zweckmäßig sein, wenn das Analyseziel mit diesem Verfahren vollständig erreicht wird. Liegen der Untersuchung hingegen mehrere unterschiedliche Aufgabenstellungen zugrunde, d.h. im Rahmen der Datenauswertung sollen eine Reihe von Analysemethoden angewandt werden, dann stellt die Bestimmung von Imputationswerten für die fehlenden Daten eine unter Umständen geeignetere Lösung dar. Des weiteren sind die Vor- und Nachteile der einzelnen Strategien in Betracht zu ziehen. Diese sind in der nachfolgenden Tabelle 5.2 zusammenfassend dargestellt:

Strategie	**Vorteile**	**Nachteile**
Eliminierung	Einfache Anwendbarkeit, Auswertung auf Basis vollständiger Daten	Zum Teil erheblicher Informationsverlust, im allgemeinen werden Objekte bzw. Merkmale aus der weiteren Analyse ausgeschlossen
Imputation	Kein Informationsverlust, Auswertung auf Basis einer vollständigen Daten- bzw. Distanzmatrix	Mögliche Verzerrung der Analyseergebnisse aufgrund der Verwendung von Schätzwerten für die fehlenden Daten
Parameterschätzung	Unverzerrte Schätzer für die jeweiligen Parameter	Nur die auf den ermittelten Parametern basierenden Auswertungsverfahren können herangezogen werden
Anwendung multivariater Analyseverfahren	Unmittelbares Erreichen eines entsprechenden Analyseziels	Mögliche Verzerrung der Analyseergebnisse aufgrund der ausschließlichen Verwendung der vorhandenen Daten
Sensitivitätsbetrachtung	Berücksichtigung möglicher Auswirkungen von MD-Verfahren auf die Imputationswerte bzw. Analyseergebnisse	Zeitaufwendig, im allgemeinen resultieren eine Vielzahl unterschiedlicher Imputationswerte bzw. Analyseergebnisse

Tabelle 5.2: Vor- und Nachteile der MD-Strategien

Die Auswahl einer geeigneten Strategie bzw. eines geeigneten MD-Verfahrens erfolgt also insgesamt unter Berücksichtigung der an das Datenmaterial gestellten Anforderungen, der Zielsetzung der Analyse sowie der Vor- und Nachteile der einzelnen Strategien.

Da die Durchführung einer Datenauswertung in der heutigen Zeit ohne die Verwendung der EDV nicht mehr vorstellbar ist, ist die Verfügbarkeit der in dieser Arbeit vorgestellten Methoden in entsprechenden Softwarepaketen von entscheidender Bedeutung. Spe-

zielle Softwarelösungen existieren bislang nur im wissenschaftlichen Bereich (vgl. z.B. *Berger, 1979*, *Engelman, 1982*, *Brouwer, Vijn, 1980*, *Schwab, 1991*) und sind im allgemeinen nicht zugänglich. Darüber hinaus handelt es sich bei diesen speziellen Programmen meist nur um die Umsetzung eines einzigen MD-Verfahrens auf den PC. In den kommerziell vertriebenen Statistiksoftwarepaketen wird die Möglichkeit fehlender Werte zwar im allgemeinen zugelassen, jedoch sind diese Programme im Hinblick auf die vorhandenen Methoden zur Analyse und Behandlung der fehlenden Daten meist stark eingeschränkt. Darüber hinaus gehen diese Programmpakete ausnahmslos von einer unvollständigen Datenmatrix aus. Die Tabelle 5.3 gibt einen Überblick der verfügbaren Methoden zur Strukturanalyse sowie der vorhandenen Verfahren zur Behandlung der fehlenden Daten in einigen Standardstatistiksoftwarepaketen.

Softwarepaket	**Methoden zur Strukturanalyse**	**MD-Verfahren**
BMDP PC-90	Missing-Data-Maße, grafische Darstellung der MD-Muster	Complete-case analysis, available-case analysis, Mittelwertimputation, Regressionsimputation, EM-Algorithmus
CSS Statistica 3.1		Complete-case analysis, available-case analysis, Mittelwertimputation
MICROSTAT II		Complete-case analysis, Regressionsimputation
NCSS 5.03	Missing-Data-Maße, grafische Darstellung der MD-Muster	Complete-case analysis, available-case analysis, Mittelwertimputation, Regressionsimputation
SPSS 5.01		Complete-case analysis, available-case analysis
STATA 3.0		Complete-case analysis, Regressionsimputation
STATGRAPHICS Plus 5.2		Complete-case analysis, available-case analysis
SYSTAT 5.21		Complete-case analysis, available-case analysis

Tabelle 5.3: Verfügbare MD-Methoden einiger Statistiksoftwarepakete

Weitere Verfahren können zwar meist mit der zur Verfügung stehenden Prozeduralsprache oder über Makros implementiert werden, jedoch erwartet der Anwender in einem umfassenden Softwarepaket auch die Möglichkeit einer umfassenden Behandlung

fehlender Daten, zumal er im allgemeinen an schnell verfügbaren Ergebnissen interessiert ist. Die adäquate praktische Umsetzung der theoretisch denkbaren und in dieser Arbeit vorgestellten Methoden im Hinblick auf eine EDV-gestützte Analyse einer unvollständigen Daten- oder Distanzmatrix sollte in der zukünftigen Forschungsarbeit daher im Vordergrund stehen.

Anhang A: Datenmatrix von 15 Statistiksoftwarepaketen für den PC

Merkmale / Objekte	Preisniveau	Benutzeroberfläche	Programmierbarkeit	Deskriptive Statistik	Testverfahren	Multivariate Verfahren	Spezialgebiete	Businessgrafiken	Statistische Grafiken
BMDP	hoch	M/K	ja	92.66	56.00	59.28	48.66	0.00	20.00
CRUNCH	mittel	M/K	nein	85.33	62.00	15.71	0.00	20.00	11.25
CSS	mittel	M/K	ja	80.33	64.00	56.85	77.00	100.00	83.00
MICROSTAT II	niedrig	M/K	nein	61.66	79.00	14.28	2.66	0.00	9.00
MINITAB	mittel	K	ja	62.33	64.00	14.85	31.00	0.00	28.24
NCSS	mittel	M	nein	76.66	67.00	49.00	80.33	92.00	69.75
P-STAT	hoch	M/K	ja	92.66	100.00	24.57	55.66	46.00	22.25
RS/1	hoch	M/K	ja	54.33	73.00	10.28	27.66	49.00	51.00
SAS	hoch	M/K	ja	89.66	78.00	71.14	94.00	100.00	51.25
SPSS	hoch	M/K	ja	89.33	86.00	60.85	52.00	86.00	20.75
STATA	mittel	M/K	ja	65.66	68.00	27.42	42.66	89.00	36.25
STATGRAPHICS	gehoben	M/K	nein	86.66	78.00	33.71	56.33	97.00	42.00
STATISTIX	niedrig	M	nein	67.33	81.00	15.14	22.00	0.00	16.00
STATPAC GOLD	mittel	M/K	nein	77.66	68.00	17.71	39.66	43.00	21.25
SYSTAT	gehoben	M/K	ja	86.66	64.00	61.00	45.00	89.00	85.25

Legende:
M = Menüsteuerung
K = Kommandosteuerung
M/K = Kombination aus Menü- und Kommandosteuerung

Tabelle A.1: Datenmatrix von 15 Statistiksoftwarepaketen für den PC

Erläuterungen:

Die in Tabelle A.1 dargestellte Datenmatrix gibt auszugsweise die Datengrundlage einer vom Institut für Statistik und Mathematische Wirtschaftstheorie der Universität Augs-

burg durchgeführten Marktstudie von Statistiksoftware für den PC wieder. Bei den Merkmalen **Deskriptive Statistik**, **Testverfahren**, **Multivariate Verfahren**, **Spezialgebiete**, **Businessgrafiken** sowie **Statistische Grafiken** handelt es sich jeweils um aggregierte Scoringwerte, die der prozentualen Leistungsbreite und -tiefe der Softwarepakete bezüglich aller betrachteten Eigenschaften und Fähigkeiten in den einzelnen Bereichen entsprechen. Damit können diese Merkmale als kardinal skaliert betrachtet werden. Die anderen Merkmale besitzen die folgenden Skalenniveaus: **Preis** - ordinal, **Benutzeroberfläche** - nominal polytom, **Programmierbarkeit** - nominal dichotom. Für weitere Ausführungen sei auf *Bausch und Bankhofer (1992, S. 285-292, 300-301)* verwiesen.

Um diese Datenmatrix in den Beispielen dieser Arbeit anwenden zu können, sollen die folgenden vier Fälle fehlender Merkmalsausprägungen unterschieden werden. Dabei wurde grundsätzlich berücksichtigt, daß der Anteil der fehlenden Daten insgesamt höchstens 10 Prozent beträgt.

Fall 1: Die Auswahl der als fehlend anzusehenden Merkmalsausprägungen erfolgt zufällig. Dabei wurden die folgenden Aspekte berücksichtigt:

- In jedem vorhanden Skalentyp sollen fehlende Daten vorliegen.
- Bei drei zufällig ausgewählten Merkmalen sollen keine Daten fehlen.

Damit ergab sich die folgende Vorgehensweise bei der Auswahl der als fehlend anzusehenden Merkmalsausprägungen:

1. Die Merkmale wurden der Reihe nach von 1 bis 9 numeriert.
2. Mittels eines Zufallszahlengenerators wurden dann drei unterschiedliche, im Intervall $[3.5,9.5\rangle$ gleichverteilte und anschließend ganzzahlig gerundete Zufallszahlen gezogen, um so die Merkmale zu bestimmen, die keine fehlenden Werte enthalten sollen.
3. Die Merkmalsausprägungen der verbleibenden Datenmatrix mit sechs Merkmalen und 15 Objekten wurden spaltenweise der Reihe nach von 1 bis 90 numeriert.
4. Mittels eines Zufallszahlengenerators wurden dann 13 unterschiedliche, im Intervall $[0.5,90.5\rangle$ gleichverteilte und anschließend ganzzahlig gerundete Zufallszahlen gezogen, um so schließlich die als fehlend anzusehenden Merkmalsausprägungen zu bestimmen.

Die auf diese Art ermittelten Werte sind in der Datenmatrix der Tabelle A.1 umrahmt.

Fall 2: Die Auswahl der als fehlend anzusehenden Merkmalsausprägungen erfolgt systematisch in der Art, daß eine denkbare Abhängigkeit des Fehlens der Daten von den tatsächlichen Realisierungen dieser Werte berücksichtigt wird. Dazu sollen die fünf höchsten Ausprägungen des Merkmals Multivariate Verfahren, die in der Datenmatrix der Tabelle A.1 durch eine Schattierung der entsprechenden Felder kenntlich gemacht sind, als fehlend betrachtet werden.

Fall 3: Die Auswahl der als fehlend anzusehenden Merkmalsausprägungen erfolgt systematisch in der Art, daß eine denkbare Abhängigkeit des Fehlens der Daten vom Fehlen anderer Merkmalsausprägungen berücksichtigt wird. Um eine entsprechende, wechselseitige Abhängigkeit beispielhaft darzustellen, sollen bei den Merkmalen Businessgrafiken und Statistische Grafiken für die ersten fünf Objekte beide Merkmalsausprägungen als fehlend betrachtet werden. Die auf diese Art festgelegten Werte sind durch eine Schattierung der entsprechenden Felder in der Datenmatrix der Tabelle A.1 gekennzeichnet.

Fall 4: Die Auswahl der als fehlend anzusehenden Merkmalsausprägungen erfolgt systematisch in der Art, daß eine denkbare Abhängigkeit des Fehlens der Daten von den Werten anderer Merkmalsausprägungen berücksichtigt wird. Dazu sollen für die fünf Objekte, deren Ausprägungen beim Merkmal Deskriptive Statistik die niedrigsten Werte annehmen, die Daten beim Merkmal Preis als fehlend betrachtet werden. Die auf diese Art festgelegten Werte sind durch eine Schattierung der entsprechenden Felder in der Datenmatrix der Tabelle A.1 gekennzeichnet.

Anhang B: Distanzmatrix von 10 Mittelklasseautomobilen

Objekte / Objekte	Audi 90	BMW 3er	Ford Sierra	Mazda 626	Nissan Primera	Opel Vectra	Peugeot 405	Renault 21	Toyota Carina	VW Passat
Audi 90	0.00	2.57	7.88	5.15	4.95	3.95	4.25	4.65	5.23	3.95
BMW 3er	2.57	0.00	9.43	5.32	5.33	4.45	5.82	4.76	5.74	5.83
Ford Sierra	7.88	9.43	0.00	5.65	2.58	5.48	4.32	3.55	4.32	6.35
Mazda 626	5.15	5.32	5.65	0.00	2.35	3.95	2.65	3.45	2.08	4.28
Nissan Primera	4.95	5.33	2.58	2.35	0.00	4.96	2.06	3.58	2.68	4.12
Opel Vectra	3.95	4.45	5.48	3.95	4.96	0.00	3.02	3.85	4.85	3.25
Peugeot 405	4.25	5.82	4.32	2.65	2.06	3.02	0.00	2.05	7.18	1.92
Renault 21	4.65	4.76	3.55	3.45	3.58	3.85	2.05	0.00	5.23	1.78
Toyota Carina	5.23	5.74	4.32	2.08	2.68	4.85	7.18	5.23	0.00	6.56
VW Passat	3.95	5.83	6.35	4.28	4.12	3.25	1.92	1.78	6.56	0.00

Tabelle B.1: Distanzmatrix von 10 Mittelklasseautomobilen

Erläuterungen:

Die in Tabelle B.1 dargestellte Distanzmatrix enthält die Mittelwerte der bei 10 Personen erhobenen paarweisen Distanzen von 10 Mittelklasseautomobilen. Jede Person sollte dabei die subjektiv empfundene, optische Ähnlichkeit zwischen jeweils zwei Fahrzeugen durch einen Wert aus dem Intervall [0,10] angeben, wobei der Wert 0 die Identität und der Wert 10 eine maximal mögliche Verschiedenheit innerhalb eines Fahrzeugpaares zum Ausdruck bringt. Um eine Vergleichbarkeit der Skalen der 10 Personen und damit eine sinnvolle Mittelwertbildung zu gewährleisten, sollte jede Person den Wert 10 für mindestens eines der 45 unterschiedlichen Paare vergeben.

Um diese Distanzmatrix in den Beispielen dieser Arbeit anwenden zu können, sollen die folgenden drei Fälle fehlender paarweiser Distanzen unterschieden werden. Dabei wur-

de grundsätzlich berücksichtigt, daß der Anteil der fehlenden paarweisen Distanzen insgesamt höchstens 20 Prozent beträgt.

Fall 1: Die Auswahl der als fehlend anzusehenden paarweisen Distanzen erfolgt zufällig. Dabei ergab sich die folgende Vorgehensweise:

1. Die relevanten paarweisen Distanzen der oberen Dreiecksmatrix wurden zeilenweise der Reihe nach von 1 bis 45 numeriert.
2. Mittels eines Zufallszahlengenerators wurden dann neun unterschiedliche, im Intervall $[0.5, 45.5\rangle$ gleichverteilte und anschließend ganzzahlig gerundete Zufallszahlen gezogen, um so die als fehlend anzusehenden paarweisen Distanzen zu bestimmen.

Die auf diese Art ermittelten Werte sind in der oberen Dreiecksmatrix der Tabelle B.1 umrahmt.

Fall 2: Die Auswahl der als fehlend anzusehenden paarweisen Distanzen erfolgt systematisch in der Art, daß eine denkbare Abhängigkeit des Fehlens der paarweisen Distanzen von den tatsächlichen Realisierungen dieser Werte berücksichtigt wird. Dazu sollen die neun höchsten Distanzwerte, die in der unteren Dreiecksmatrix der Tabelle B.1 durch eine Schattierung der entsprechenden Felder kenntlich gemacht sind, als fehlend betrachtet werden.

Fall 3: Die Auswahl der als fehlend anzusehenden paarweisen Distanzen erfolgt systematisch in der Art, daß eine denkbare Abhängigkeit des Fehlens der paarweisen Distanzen vom Fehlen anderer Distanzwerte berücksichtigt wird. Um eine entsprechende, wechselseitige Abhängigkeit beispielhaft darzustellen, sollen alle paarweisen Distanzen zwischen den inländischen Fahrzeugen als fehlend betrachtet werden. Die auf diese Art festgelegten Werte sind durch eine Schattierung der entsprechenden Felder in der oberen Dreiecksmatrix der Tabelle B.1 gekennzeichnet.

Symbolverzeichnis

Symbol	Bedeutung
$N = \{1,...,n\}$	Objektmenge
$M = \{k_1,...,k_m\}$	Merkmalsmenge
a_{ik}	Ausprägung des Merkmals k bei Objekt i
$A = (a_{ik})_{n,m} = \begin{pmatrix} a_{11} & \dots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nm} \end{pmatrix}$	Datenmatrix mit n Zeilen (Objekten) und m Spalten (Merkmalen)
A^{obs}	vorhandener Teil der Ausprägungen aus A
A^{mis}	fehlender Teil der Ausprägungen aus A
p	Anzahl der Objekte ohne fehlende Daten
q	Anzahl der Merkmale ohne fehlende Daten
$A_{obs} = (a_{ik})_{p,m}$ bzw. $A_{obs} = (a_{ik})_{n,q}$	vollständige Teilmatrix von A
$A_{mis} = (a_{ik})_{n-p,m}$ bzw. $A_{mis} = (a_{ik})_{n,m-q}$	unvollständige Teilmatrix von A
$a_k = \begin{pmatrix} a_{1k} \\ \vdots \\ a_{nk} \end{pmatrix}$	Merkmalsvektor k (Vektor der Ausprägungen des Merkmals k bei allen Objekten)
$a^i = \begin{pmatrix} a_{i1} \\ \vdots \\ a_{im} \end{pmatrix}$	Objektvektor i (Vektor der Merkmalsausprägungen des Objekts i)
$\overline{a}_k$	Mittelwert des Merkmals k
r_{kl}	Korrelationskoeffizient der Merkmale k und l
$R = (r_{kl})_{m,m} = \begin{pmatrix} r_{11} & \dots & r_{1m} \\ \vdots & & \vdots \\ r_{m1} & \dots & r_{mm} \end{pmatrix}$	Korrelationsmatrix

Symbol	Bedeutung
s_{kk}	Varianz des Merkmals k
s_{kl}	Kovarianz der Merkmale k und l
$S = (s_{kl})_{m,m} = \begin{pmatrix} s_{11} & \dots & s_{1m} \\ \vdots & & \vdots \\ s_{m1} & \dots & s_{mm} \end{pmatrix}$	Kovarianzmatrix
$\tilde{A} = (\tilde{a}_{ik})_{n,m} = \begin{pmatrix} \tilde{a}_{11} & \dots & \tilde{a}_{1m} \\ \vdots & & \vdots \\ \tilde{a}_{n1} & \dots & \tilde{a}_{nm} \end{pmatrix}$	standardisierte und vervollständigte Datenmatrix mit $\tilde{a}_{ik} = \begin{cases} \frac{a_{ik} - \bar{a}_k}{\sqrt{s_{kk}}} & \text{falls } a_{ik} \text{ vorhanden} \\ 0 & \text{sonst} \end{cases}$
$\tilde{A}_{obs} = (\tilde{a}_{ik})_{n,q}$	Teilmatrix von $\tilde{A}$ mit den Ausprägungen $\tilde{a}_{ik}$ für die q Merkmale ohne fehlende Daten
$\tilde{A}_{mis} = (\tilde{a}_{ik})_{n,m-q}$	Teilmatrix von $\tilde{A}$ mit den Ausprägungen $\tilde{a}_{ik}$ für die $(m-q)$ Merkmale mit fehlenden Daten
$V = (v_{ik})_{n,m} = \begin{pmatrix} v_{11} & \dots & v_{1m} \\ \vdots & & \vdots \\ v_{n1} & \dots & v_{nm} \end{pmatrix}$	Indikatormatrix der fehlenden Ausprägungen mit $v_{ik} = \begin{cases} 1 & \text{falls } a_{ik} \text{ vorhanden} \\ 0 & \text{sonst} \end{cases}$
$v_k = \begin{pmatrix} v_{1k} \\ \vdots \\ v_{nk} \end{pmatrix}$	Merkmalsvektor k der Indikatormatrix V
$v^i = \begin{pmatrix} v_{i1} \\ \vdots \\ v_{im} \end{pmatrix}$	Objektvektor i der Indikatormatrix V
$v_{i\bullet}^{ind} = \begin{cases} 1 & \text{falls } v_{ik} = 1 \quad \forall k \in M \\ 0 & \text{sonst} \end{cases}$	MD-Indikator für Objekt i
$v_{\bullet k}^{ind} = \begin{cases} 1 & \text{falls } v_{ik} = 1 \quad \forall i \in N \\ 0 & \text{sonst} \end{cases}$	MD-Indikator für Merkmal k

Symbol	Bedeutung
$v_{i\bullet}^{mis}$	Anzahl der fehlenden Daten bei Objekt i
$v_{\bullet k}^{mis}$	Anzahl der fehlenden Daten bei Merkmal k
$v_{i\bullet}^{obs}$	Anzahl der vorhandenen Daten bei Objekt i
$v_{\bullet k}^{obs}$	Anzahl der vorhandenen Daten bei Merkmal k
v^{mis}	Anzahl der fehlenden Daten in der Datenmatrix
v^{obs}	Anzahl der vorhandenen Daten in der Datenmatrix
$\tilde{v}_{i\bullet}^{mis}$	Anteil der fehlenden Daten bei Objekt i
$\tilde{v}_{\bullet k}^{mis}$	Anteil der fehlenden Daten bei Merkmal k
$\tilde{v}_{i\bullet}^{obs}$	Anteil der vorhandenen Daten bei Objekt i
$\tilde{v}_{\bullet k}^{obs}$	Anteil der vorhandenen Daten bei Merkmal k
$\tilde{v}^{mis}$	Anteil der fehlenden Daten in der Datenmatrix
$\tilde{v}^{obs}$	Anteil der vorhandenen Daten in der Datenmatrix
$N_{kl} = \{i: v_{ik} = 1 \wedge v_{il} = 1\}$	Menge von Objekten, deren Ausprägungen bezüglich der Merkmale k und l vorhanden sind
$N_k = \{i: v_{ik} = 1\}$	Menge von Objekten, deren Ausprägungen bezüglich des Merkmals k vorhanden sind
$M_{ij} = \{k: v_{ik} = 1 \wedge v_{jk} = 1\}$	Menge von Merkmalen, deren Ausprägungen bei den Objekten i und j vorhanden sind
$M_i = \{k: v_{ik} = 1\}$	Menge von Merkmalen, deren Ausprägungen bei Objekt i vorhanden sind

Symbol	Bedeutung
d_{ij}	Distanzindex der Objekte i und j
$D = (d_{ij})_{n,n} = \begin{pmatrix} d_{11} & \dots & d_{1n} \\ \vdots & & \vdots \\ d_{n1} & \dots & d_{nn} \end{pmatrix}$	Distanzmatrix der n Objekte
D^{obs}	vorhandener Teil der Distanzen aus D
D^{mis}	fehlender Teil der Distanzen aus D
D_{obs}	vollständige Teilmatrix von D
D_{mis}	unvollständige Teilmatrix von D
o	Anzahl der Objekte ohne fehlende Distanzen
$W = (w_{ij})_{n,n} = \begin{pmatrix} w_{11} & \dots & w_{1n} \\ \vdots & & \vdots \\ w_{n1} & \dots & w_{nn} \end{pmatrix}$	Indikatormatrix der fehlenden Distanzen mit $w_{ij} = \begin{cases} 1 & \text{falls } d_{ij} \text{ vorhanden} \\ 0 & \text{sonst} \end{cases}$
$w_i^{ind} = \begin{cases} 1 & \text{falls } w_{ij} = 1 \quad \forall j \in N, j \neq i \\ 0 & \text{sonst} \end{cases}$	MD-Indikator für Objekt i
w_i^{mis}	Anzahl der fehlenden Distanzen bei Objekt i
w_i^{obs}	Anzahl der vorhandenen Distanzen bei Objekt i
w^{mis}	Anzahl der fehlenden Distanzen in der Distanzmatrix
w^{obs}	Anzahl der vorhandenen Distanzen in der Distanzmatrix
$\tilde{w}_i^{mis}$	Anteil der fehlenden Distanzen bei Objekt i
$\tilde{w}_i^{obs}$	Anteil der vorhandenen Distanzen bei Objekt i
$\tilde{w}^{mis}$	Anteil der fehlenden Distanzen in der Distanzmatrix

Symbol	Bedeutung
$\tilde{w}^{obs}$	Anteil der vorhandenen Distanzen in der Distanzmatrix
$N^{ij} = \{h: w_{ih} = 1 \wedge w_{jh} = 1, h \neq i,j\}$	Menge von Objekten, deren Distanzen bezüglich der Objekte i und j vorhanden sind
$N^{i} = \{h: w_{ih} = 1, h \neq i\}$	Menge von Objekten, deren Distanzen bezüglich des Objekts i vorhanden sind
$f(x)$	Dichte- bzw. Wahrscheinlichkeitsfunktion
$f(x\|y)$	bedingte Dichte- bzw. Wahrscheinlichkeitsfunktion
P	Wahrscheinlichkeitsmaß
$N(\mu;\sigma)$	Normalverteilung
$\Phi(x)$	Verteilungsfunktion der Standardnormalverteilung
$E(X)$	Erwartungswert der Zufallsvariablen X
$Cov(X,Y)$	Kovarianz der Zufallsvariablen X und Y
$\lambda_1,\ldots,\lambda_m$	Eigenwerte einer $(m \times m)$-Matrix
$D_\lambda = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \lambda_m \end{pmatrix}$	Diagonalmatrix der Eigenwerte
$u_q = \begin{pmatrix} u_{1q} \\ \vdots \\ u_{mq} \end{pmatrix}$	Eigenvektor zum Eigenwert λ_q
$U = (u_1,\ldots,u_m) = \begin{pmatrix} u_{11} & \cdots & u_{1m} \\ \vdots & & \vdots \\ u_{m1} & \cdots & u_{mm} \end{pmatrix}$	Matrix der zu den Eigenwerten von D_λ zugehörigen Eigenvektoren

Symbol	Bedeutung
I	Einheitsmatrix
$\mathcal{K}=\{K_1,\ldots,K_s\}$	Klassifikation der Objekte mit s Klassen
β_k	zum Merkmal k gehörender Regressionskoeffizient
β_0	Absolutglied
x_{ij}	Beobachtungswert für die i-te Stufe des ersten Faktors und die j-te Stufe des zweiten Faktors
α_i	Effekt der i-te Stufe des ersten Faktors
β_j	Effekt der j-te Stufe des zweiten Faktors
ε_{ij}	zufälliger Fehler für den Beobachtungswert x_{ij}
g_k	zum Merkmal k gehörender Diskriminanzkoeffizient
$X=(x_{ik})_{n,m}=\begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix}$	Faktorwertematrix mit den Faktorwerten x_{ik}
$F=(f_{kl})_{m,m}=\begin{pmatrix} f_{11} & \cdots & f_{1m} \\ \vdots & & \vdots \\ f_{m1} & \cdots & f_{mm} \end{pmatrix}$	Ladungsmatrix mit den Faktorladungen f_{kl}
$F_{obs}=(f_{kl})_{q,m}$	Teilmatrix von F mit den Faktorladungen für die q Merkmale ohne fehlende Daten
$F_{mis}=(f_{kl})_{m-q,m}$	Teilmatrix von F mit den Faktorladungen für die $(m-q)$ Merkmale mit fehlenden Daten
$[a;b]$	abgeschlossenes Intervall zwischen a und b
$\langle a;b\rangle$	offenes Intervall zwischen a und b
$[a;b\rangle, \langle a;b]$	halboffene Intervalle zwischen a und b

Abkürzungsverzeichnis

Abkürzung	Bedeutung
AID	automatic interaction detector
ANCOVA	analysis of covariance
ANOVA	analysis of variance
EM	expectation maximization
MAR	missing at random
MARC	missing at random within classes
MCAR	missing completely at random
MCARC	missing completely at random within classes
MD	missing data
MVL	missing value linkage
OAR	observed at random
PAC	pyramidal ascending clustering
PACII	pyramidal ascending clustering with incomplete information
SVD	singular value decomposition
RP	Ridge-Prozedur

Literaturverzeichnis

Afifi, A.A.; Elashoff, R.M. (1966): Missing Observations in Multivariate Statistics. I: Review of the Literature, *Journal of the American Statistical Association, 61, S. 595-604*

Afifi, A.A.; Elashoff, R.M. (1967): Missing Observations in Multivariate Statistics. II: Point Estimation in Simple Linear Regression, *Journal of the American Statistical Association, 62, S. 10-29*

Albrecht, P. (1980): On the Correct Use of the Chi-square Goodness-of-fit-Test, *Scandinavian Actuarial Journal, S. 149-160*

Ambrosi, K. (1981): Distanzen und Präordnungen bei qualitativen Merkmalen, Arbeitspapiere zur Mathematischen Wirtschaftsforschung, Heft 47/1981, Universität Augsburg

Anderson, A.B.; Basilevsky, A.; Hum, D.J. (1983): Missing Data: A Review of the Literature, *Rossi, P.H.; Wright, J.D.; Anderson, A.B. (Hrsg.), Handbook of Survey Research, New York, S. 415-493*

Anderson, T.W. (1957): Maximum Likelihood Estimates for a Multivariate Normal Distribution when some Observations are Missing, *Journal of the American Statistical Association, 52, S. 200-203*

Backhaus, K.; Erichson, B.; Plinke, W.; Weiber, R. (1990): Multivariate Analysemethoden: eine anwendungsorientierte Einführung, 6. Auflage, Springer, Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong, Barcelona

Baker, S.G.; Laird, N.M. (1988): Regression Analysis for Categorical Variables with Outcome Subjects to Nonignorible Nonresponse, *Journal of the American Statistical Association, 83, S. 62-69*

Bamberg, G.; Baur, F. (1993): Statistik, 8. Auflage, Oldenbourg, München, Wien

Bamberg, G.; Schittko, U.K. (1979): Einführung in die Ökonometrie, Fischer, Stuttgart, New York

Barlett, M.S. (1937): Some Examples of Statistical Methods of Research in Agriculture and Applied Botany, *Journal of the Royal Statistical Society, 4, Series B, S. 137-170*

Bausch, T.; Bankhofer, U. (1992): Statistical Software Packages for PCs - A Market Survey, *Statistical Papers / Statistische Hefte, 33, S. 283-306*

Bausch, T.; Opitz, O. (1993): PC-gestützte Datenanalyse mit Fallstudien aus der Marktforschung, Vahlen, München

Beale, E.M.L.; Little, R.J.A. (1975): Missing Values in Multivariate Analysis, *Journal of the Royal Statistical Society, 37, Series B, S. 129-145*

Berger, M.P.F. (1979): A FORTRAN IV Program for the Estimation of Missing Data, *Behavior Research Methods and Instrumentation, 11/3, S. 395-396*

Blumenthal, S. (1968): Multinomial Sampling with Partially Categorized Data, *Journal of the American Statistical Association, 63, S. 542-551*

Bock, H.H. (1974): Automatische Klassifikation, Vandenhoeck & Ruprecht, Göttingen

Brouwer, U.; Vijn, P. (1980): A Program to Estimate the Correlation Coefficient in Incomplete Datasets, *International Association for Statistical Computing: COMPSTAT 1980, Wien, S. 194-200*

Brown, C.H. (1983): Asymptotic Comparison of Missing Data Procedures for Estimating Factor Loadings, *Psychometrika, 48, S. 269-291*

Buck, S.F. (1960): A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer, *Journal of the Royal Statistical Society, 22, Series B, S. 302-306*

Büning, H.; Trenkler, G. (1978): Nichtparametrische statistische Methoden, de Gruyter, Berlin, New York

Cailliez, F. (1983): The Analytical Solution of the Additive Constant Problem, *Psychometrika, 48, S. 305-308*

Chan, L.S.; Dunn, O.J. (1972): The Treatment of Missing Values in Discriminant Analysis - I. The Sampling Experiment, *Journal of the American Statistical Association, 67, S. 473-477*

Chan, L.S.; Dunn, O.J. (1974): A Note on the Asymptotic Aspect of the Treatment of Missing Values in Discriminant Analysis, *Journal of the American Statistical Association, 69, S. 672-673*

Chapman. D.W. (1976): A Survey of Nonresponse Imputation Procedures, *Journal of the American Statistical Association, Proceedings of the Social Statistical Section, S. 245-251*

Chapman, D.W. (1983a): An Investigation of Nonresponse Imputation Procedures for the Health and Nutrition Examination Survey, *Madow, W.G.; Nisselson, H.; Olkin, I. (Hrsg.), Incomplete Data in Sample Surveys, 1, Report and Case Studies, Academic Press, New York, S. 435-483*

Chapman, D.W. (1983b): The Impact of Substitution on Survey Estimates, *Madow, W.G.; Nisselson, H.; Olkin, I., Incomplete Data in Sample Surveys, 2, Theory and Bibliographies, Academic Press, New York, S. 45-61*

Chen, T.; Fienberg, S.E. (1974): Two-dimensional Contingency Tables with Both Completely and Partially Cross-Classified Data, *Biometrics, 30, S. 629-642*

Chen, T.; Fienberg, S.E. (1976): The Analysis of Contingency Tables with Incompletely Classified Data, *Biometrics, 32, S. 132-144*

Cohen, J.; Cohen, P. (1975): Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, Hillsdale

Davison, M.L. (1983): Multidimensional Scaling, Wiley, New York

Dear, R.E. (1959): A Principal Component Missing Data Method for Multiple Regresion Models, Technical Report SP-86, System Development Corporation, Santa Monica

Dempster, A.P. (1971): An Overview of Multivariate Data Analysis, *Journal of Multivariate Analysis, 1, S. 316-346*

Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977): Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society, 39, Series B, S. 1-38*

De Soete, G. (1984a): Ultrametric Tree Representations of Incomplete Dissimilitarity Data, *Journal of Classification, 1, S. 235-242*

De Soete, G. (1984b): Additive Tree Representations of Incomplete Dissimilitarity Data, *Quality and Quantity, 18, S. 387-393*

Edgett, G.L. (1956): Multiple Regression with Missing Observations among the Independent Variables, *Journal of the American Statistical Association, 51, S. 122-132*

Engelman, L. (1982): An Efficient Algorithm for Computing Covariance Matrices from Data with Missing Values, *Communications in Statistics, Simulation and Computation, B11/1, S. 113-121*

Evans, R.W.; Cooley, P.C; Piserchia, P.V. (1979): A Test for Evaluating Missing Data Imputation Procedures, *ASA Proceedings of the Social Statistics Section, S. 469-474*

Federspiel, C.F.; Monroe, R.J.; Greenberg, B.G. (1959): An Investigation of Some Multiple Regression Methods for Incomplete Samples, University of North Carolina, Institute of Statistics, Memeo Series, No. 236

Fellegi, I.P.; Holt, D. (1976): A Systematic Approach to Automatic Edit and Imputation, *Journal of the American Statistical Association, 71, S. 17-35*

Fienberg, S.E. (1972): The Analysis of Incomplete Multiway Contingency Tables, *Biometrics, 28*

Ford, B.L. (1976): Missing Data Procedures: A Comparative Study, *ASA Proceedings of the Social Statistics Section, S. 324-329*

Ford, B.L. (1983): An Overview of Hot-Deck Procedures, *Madow, W.G.; Nisselson, H.; Olkin, I. (Hrsg.), Incomplete Data in Sample Surveys, 2, Theory and Bibliographies, Academic Press, New York, S. 185-207*

Frane, J.W. (1976): Some Simple Procedures for Handling Missing Data in Multivariate Analysis, *Psychometrika, 41, S. 409-415*

Frane, J.W. (1978): Missing Data and BMDP: Some Pragmatic Approaches, *ASA Proceedings of the Statistical Computing Section, S. 27-33*

Freund, R.J.; Minton, P.D. (1979): Regression Methods, Dekker, New York

Fuchs, C. (1982): Maximum Likelihood Estimation and Model Selection in Contigency Tables with Missing Data, *Journal of the American Statistical Association, 77, S. 270-278*

Furnas, G.W. (1989): Metric Family Portraits, *Journal of Classification, 6, S. 7-52*

Glasser, M. (1964): Linear Regression Analysis with Missing Observations Among the Independent Variables, *Journal of the American Statistical Association, 59, S. 834-844*

Gleason, T.C.; Staelin, R.A. (1975): A Proposal for Handling Missing Data, *Psychometrika, 40, S. 229-252*

Gower, J.C. (1971): A General Coefficient of Similarity and Some of Its Properities, *Biometrics, 27, S. 857-872*

Greenless, W.S.; Reece, J.S.; Zieschang, K.D. (1982): Imputation of Missing Values when the Probability of Response Depends on the Variables Being Imputed, *Journal of the American Statistical Association, 77, S. 251-261*

Haitovsky, Y. (1968): Missing Data in Regression Analysis, *Journal of the Royal Statistical Society, 30, Series B, S. 67-82*

Hartley, H.O.; Hocking, R.R. (1971): The Analysis of Incomplete Data, *Biometrics, 27, S. 783-823*

Hartung, J. (1989): Statistik, Lehr- und Handbuch der angewandten Statistik, 7. Auflage, Oldenbourg, München, Wien

Hartung, J.; Elpelt, B. (1992): Multivariate Statistik, Lehr und Handbuch der angewandten Statistik, 4. Auflage, Oldenbourg, München, Wien

Hauke, W. (1992): Darstellung struktureller Zusammenhänge und Entwicklungen in Input-Output-Tabellen, Eul, Bergisch Gladbach, Köln

Heiberger, R.M. (1977): Regression with the Pairwise-Present Covariance Matrix: A Dangerous Practice, *ASA Proceedings of the Statistical Computing Section, S. 38-47*

Herzog, T.N.; Rubin, D.B. (1983): Using Multiple Imputations to Handle Nonresponse in Sample Surveys, *Madow, W.G.; Nisselson, H.; Olkin, I. (Hrsg.), Incomplete Data in Sample Surveys, 2, Theory and Bibliographies, Academic Press, New York, S. 209-245*

Hill, M.; Dixon, W.J. (1981): Missing Data: Search for Patterns, *ASA Proceedings of the Statistical Computing Section, S. 57-60*

Hocking, R.R. (1983): The Design and Analysis of Sample Survey with Incomplete Data: Reduction of Respondent Burden, *Madow, W.G.; Nisselson, H.; Olkin, I. (Hrsg.), Incomplete Data in sample Surveys, 3, Proceedings of the Symposium, Academic Press, New York, S. 107-124*

Hocking, R.R.; Marx, D.L. (1979): Estimating with Incomplete Data: An Improved Computational Method and the Analysis of Nested Data, *Communications in Statistical Theoretical Methodology, A8, S. 1155-1181*

Hocking, R.R.; Oxspring, H.H. (1971): Maximum Likelihood Estimation with Incomplete Multinomial Data, *Journal of the American Statistical Association, 66, S. 65-70*

Jackson, E.C. (1968): Missing Values in Linear Multiple Discriminant Analysis, *Biometrics, 24, S. 835-844*

Jobson, J.D. (1991): Applied Multivariate Data Analysis, Volume I: Regression and Experimental Design, Springer, New York, Berlin, Heidelberg

Jobson, J.D. (1992): Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods, Springer, New York, Berlin, Heidelberg

Kendall, M. (1975): Multivariate Analysis, Griffin & Company, London

Kim, J.O.; Curry, J. (1977): The Treatment of Missing Data in Multivariate Analysis, *Sociological Methods and Research, 6, S. 215-239*

Koch, G.G.; Imrey, P.B.; Reinfurt, D.W. (1972): Linear Model Analysis of Categorical Data with Incomplete Response Vectors, *Biometrics, 28, S. 663-692*

Laird, N.M. (1985): Missing Information Principle, *Kotz, S.; Johnson, N.L. (Hrsg.), Encyclopedia of Statistical Sciences, Volume 5, Wiley, New York, S. 548-552*

Little, R.J.A. (1982): Models for Nonresponse in Sample Surveys, *Journal of the American Statistical Association, 77, S. 237-250*

Little, R.J.A. (1988): A Test of Missing Completely at Random for Multivariate Data with Missing Values, *Journal of the American Statistical Association, 83, S. 1198-1202*

Little, R.J.A.; Rubin, D.B. (1987): Statistical Analysis with Missing Data, Wiley, New York

Little, R.J.A.; Schluchter, M.D. (1985): Maximum Likelihood Estimation for Mixed Continous and Categorical Data with Missing Values, *Biometrika, Nr. 72, S. 497-512*

Little, R.J.A.; Smith, P.J. (1987): Editing and Imputation for Quantitative Survey Data, *Journal of the American Statistical Association, 82, S. 58-68*

Lösel, F; Wüstendörfer, W. (1974): Zum Problem unvollständiger Datenmatrizen in der empirischen Sozialforschung, *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 26, S. 342-357*

Lord, F.M. (1955): Estimation of Parameters from Incomplete Data, *Journal of the American Statistical Association, 50, S. 870-876*

MacCallum, R.C. (1979): Recovery of Structure in Incomplete Data by ALSCAL, *Psychometrika, 44, S. 69-74*

Madow, W.G.; Nisselson, H.; Olkin, I. (Hrsg.) (1983): Incomplete Data in Sample Surveys, Volume 1, Report and Case Studies, Academic Press, New York

Malhotra, N.K.; Jain, A.K.; Pinson, C. (1988): The Robustness of MDS Configurations in the Case of Incomplete Data, *Journal of Marketing Research, 25, S. 92-102*

Matthai, A. (1951): Estimation of Parameters from Incomplete Data with Applications to Design of Sample Surveys, *Sankhya, S. 145-152*

Möntmann, V.; Bollinger, G.; Herrmann, A. (1983): Tests auf Zufälligkeit von „Missing Data", *Wilke, H. (Hrsg.), Statistiksoftware in der Sozialforschung, Berlin, S. 87-101*

Neumann, K. (1975): Operations Research Verfahren, Band I, Hanser, München, Wien

Neumann, K. (1977): Operations Research Verfahren, Band II, Hanser, München, Wien

Nordheim, E.V. (1978a): Inference from Nonrandomly Missing Data, Ph.D. Thesis, University of Minnesota

Nordheim, E.V. (1978b): Obtaining Information from Nonrandomly Missing Data, *ASA Proceedings of the Statistical Computing Section, S. 34-38*

Oh, H.L.; Scheuren, F.J. (1983): Weighting Adjustment for Unit Nonresponse, *Madow, W.G.; Nisselson, H.; Olkin, I. (Hrsg.), Incomplete Data in Sample Surveys, 2, Theory and Bibliographies, Academic Press, New York, S. 143-184*

Opitz, O. (1980): Numerische Taxonomie, UTB, Fischer, Stuttgart, New York

Opitz, O. (1995): Mathematik, Lehrbuch für Ökonomen, 5. Auflage, Oldenbourg, München, Wien

Orchard, T.; Woodbury, M.A. (1972): A Missing Information Principle: Theory and Applications, *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Nr. 1, S. 697-715*

Platek, R.; Gray, G.B. (1983): Imputation Methodology: Total Survey Error, *Madow, W.G.; Olkin, I.; Rubin, D.B. (Hrsg.), Incomplete Data in Sample Surveys, 2, Theory and Bibliographies, Academic Press, New York, S. 249-333*

Press, S.J.; Scott, A. (1974): Missing Variables in Bayesian Regression, *Fienberg, S.E.; Zellner, A. (Hrsg.), Studies in Bayesian Econometrics and Statistics, S. 259-272, North-Holland*

Rizvi, H. (1983): An Empirical Investigation of Some Item Nonresponse Adjustment Procedures, *Madow, W.G.; Nisselson, H.; Olkin, I. (Hrsg.), Incomplete Data in Sample Surveys, 1, Report and Case Studies, Academic Press, New York, S. 299-366*

Rubin, D.B. (1972): A Non-Iterative Algorithm for Least Squares Estimation of Missing Values in any Analysis of Variance Design, *Applied Statistics, 21, S. 136-141*

Rubin, D.B. (1974): Characterizing the Estimation of Parameters in Incomplete Data Problems, *Journal of the American Statistical Association, 69, S. 467-474*

Rubin, D.B. (1976): Inference and Missing Data, *Biometrika, 63, S. 581-592*

Rubin, D.B. (1977): Formalizing Subjective Notations About the Effect of Nonrespondents in Sample Surveys, *Journal of the American Statistical Association, 72, S. 538-543*

Rubin, D.B. (1978): Multiple Imputations in Sample Surveys - A Phenomenological Bayesian Approach to Nonresponse, *ASA Proceedings of the Section on Survey Research Methods, S. 20-34*

Rubin, D.B. (1979): Illustrating the Use of Multiple Imputation to Handle Nonresponse in Sample Surveys, *Bulletin of the International Statistical Institute, S. 517-532*

Rubin, D.B. (1987): Multiple Imputation for Nonresponse in Surveys, Wiley, New York

Rummel, R.J. (1970): Applied Factor Analysis, Northwestern University Press, Evanston

Sande, I.G. (1983): Hot-Deck Imputation Procedures, *Madow, W.G.; Nisselson, H.; Olkin, I. (Hrsg.), Incomplete Data in Sample Surveys, 3, Proceedings of the Symposium, Academic Press, New York, S. 339-349*

Santos, R.L. (1981): Effects of Imputation on Complex Statistics, Survey Research Center, Institute for Social Research, University of Michigan

Schader, M.; Gaul, W. (1990): Pyramidal Clustering with Missing Values, *Proceedings INRIA Conference Symbolic - Numeric Data Analysis and Learning, Nova Science, S. 523-534*

Schader, M.; Gaul, W. (1991): The MVL (Missing Values Linkage) Approach for Hierarchical Classification when Data are Incomplete, *Schader, M. (Hrsg.), Analyzing and Modeling Data and Knowledge, Proceedings of the 15th Annual Conference of the "Gesellschaft für Klassifikation e.V.", University of Salzburg, Springer, Berlin*

Schnell, R. (1985): Zur Effizienz einiger Missing-Data-Techniken - Ergebnisse einer Computer-Simulation, *ZUMA-Nachrichten 17, S. 50-74*

Schnell, R. (1986): Missing-Data Probleme in der Empirischen Sozialforschung, Dissertation, Bochum

Schwab, G. (1991): Fehlende Werte in der angewandten Statistik, Dt. Univ.-Verl., Wiesbaden

Schwertman, N.C.; Allen, D.M. (1979): Smoothing an Indefinite Variance-Covariance Matrix, *Journal of Statistical Computation and Simulation, 9, S. 183-194*

Simon, G.A.; Simonoff, J.S. (1986): Diagnostic Plots for Missing Data in Least Squares Regression, *Journal of the American Statistical Association, 81, S. 501-509*

Sodeur, W. (1974): Empirische Verfahren zur Klassifikation, Teubner, Stuttgart

Sonquist, J.A.; Dunkelberg, W.C. (1977): Survey and Opinion Research: Procedures for Processing and Analysis, Englewood Cliffs

Spence, I.; Domoney, D.W. (1974): Single Subject Incomplete Designs for Nonmetric Multidimensional Scaling, *Psychometrika, 39, S. 469-490*

Steinhausen, D.; Langer, K. (1977): Clusteranalyse: Einführung in Methoden und Verfahren der automatischen Klassifikation, de Gruyter, Berlin, New York

Stewart, D.W. (1982): Filling the Gap: A Review of the Missing Data Problem, *Walker, B.J. (Hrsg.), Assessment of Marketing Thought and Practice, Proceedings of the American Marketing Association, 48, Chicago, S. 395-399*

Takane, Y.; Young, F.W.; deLeeuw, J. (1977): Nonmetric Individual Differences Multidimensional Scaling: An Alternating Least Squares Method with Optimal Scaling Features, *Psychometrika, 42, S. 7-67*

Timm, N.H. (1970): The Estimation of Variance-Covariance and Correlation Matrices from Incomplete Data, *Psychometrika, 35, S. 417-437*

Toutenburg, H. (1992): Lineare Modelle, Physica, Heidelberg

Weber, E. (1974): Einführung in die Faktorenanalyse, Fischer, Stuttgart

Wiberg, T. (1976): Computation of Principal Components when Data are Missing, *International Association for Statistical Computing: COMPSTAT 1976, Wien, S. 229-236*

Wilkinson, G.N. (1958): Estimation of the Missing Values for the Analysis of Incomplete Data, *Biometrics, 14, S. 257-286*

Wilks, S.S. (1932): Moments and Distributions of Estimates of Population Parameters from Fragmentary Samples, *Annals of Mathematical Statistics, 2, S. 163-195*

Wishart, D. (1978): Treatment of Missing Values in Cluster Analysis, *International Association for Statistical Computing: COMPSTAT 1978, Wien, S. 281-287*

Wishart, D. (1985): Estimation of Missing Values and Diagnosis Using Hierarchical Classifications, *Computational Statistics Quarterly, 2, S. 125-134*

Wishart, D. (1986): Hierarchical Cluster Analysis with Messy Data, *Gaul, W.; Schader, M. (Hrsg.), Classification as a Tool of Research, North-Holland, S. 453-460*

Yates, F. (1933): The Analysis of Replicated Experiments When the Field Results are Incomplete, *The Empire Journal of Experimental Agriculture, 1, S. 129-142*

Stichwortverzeichnis

Neuerscheinungen des Josef Eul Verlages

REIHE: STEUER, WIRTSCHAFT UND RECHT
Herausgegeben von vBP StB Dr. Johannes Georg Bischoff, Köln; Dr. Alfred Kellermann, Vorsitzender Richter am BGH, Karlsruhe; Prof. Dr. Günter Sieben, Köln und Prof. Dr. Norbert Herzig, Köln

Band 116: Astrid Linscheidt
Methodische Grundlagen zur Ermittlung der wettbewerbsrelevanten Unternehmenssteuerbelastung in der Europäischen Union
Bergisch Gladbach 1994, 264 S., 66,— DM, ISBN 3-89012-412-7

Band 117: Frank Thiede
Ökonomische Analyse der Körperschaftsbesteuerung bei ausländischen Einkünften
Bergisch Gladbach 1994, 368 S., 76,— DM, ISBN 3-89012-413-5

Band 118: Wolfgang Sturm
Die verdeckte Gewinnausschüttung im europäischen Konzern
Bergisch Gladbach 1994, 352 S., 76,— DM, ISBN 3-89012-414-3

Band 119: Oliver Bernards
Segmentberichterstattung diversifizierter Unternehmen – Theoretische und empirische Analyse
Bergisch Gladbach 1994, 460 S., 86,— DM, ISBN 3-89012-421-6

Band 120: Joachim Wüst
Die persönliche Zurechnung der Einkünfte beim Nießbrauch
Bergisch Gladbach 1995, 276 S., 67,— DM, ISBN 3-89012-430-5

Band 121: Uwe Gotthardt
Rückstellungen und Umweltschutz
Bergisch Gladbach 1995, 400 S., 86,— DM, ISBN 3-89012-441-0

Band 122: Georg Rützel
Handels- und Steuerbilanz in Zeiten schleichender Inflation – Dargestellt am Beispiel von Deutschland, den Niederlanden und der Europäischen Gemeinschaft
Bergisch Gladbach 1995, 192 S., 64,— DM, ISBN 3-89012-442-9

Band 123: Norbert Kratz
Kapitalstrukturgestaltung im Konzern – Eine ökonomische Analyse ihrer Wohlfahrtseffekte sowie ihrer Abbildung im Rahmen der externen Rechnungslegung
Bergisch Gladbach 1995, 304 S., 75,— DM, ISBN 3-89012-444-5

Band 124: Andrea Goebel
Möglichkeiten der Entschlüsselung von Konzernkrisen mit der Methodik der integrativen Konzernabschlußanalyse – Dargestellt unter Berücksichtigung der Aussagefähigkeit der externen Rechnungslegung von Konzernen mit deutscher Muttergesellschaft und Börsennotierung an der New York Stock Exchange
Bergisch Gladbach 1995, 464 S., 89,— DM, ISBN 3-89012-450-X

Band 125: Hansjörg Flassak
Der Markt für Unternehmenskontrolle – Eine ökonomische Analyse vor dem Hintergrund des deutschen Gesellschaftsrechts
Bergisch Gladbach 1995, 420 S., 86,— DM, ISBN 3-89012-457-7

REIHE: GRÜNDUNG, INNOVATION UND BERATUNG
Herausgegeben von Prof. Dr. Dr. h. c. Norbert Szyperski, Köln; vBP StB Dr. Johannes Georg Bischoff, Köln und Dr. Heinz Klandt, Köln

Band 17: Jochen Wenz
Unternehmensgründungen aus volkswirtschaftlicher Sicht
Bergisch Gladbach 1993, 264 S., 59,— DM, ISBN 3-89012-354-6

REIHE: WISO-STUDIENTEXTE
Herausgegeben von Prof. Dr. Eckart Bomsdorf, Köln und Prof. Dr. Josef Kloock, Köln

Band 1: Eckart Bomsdorf
Deskriptive Statistik – Mit einem Anhang zur Bevölkerungs- und Erwerbsstatistik – 8., überarbeitete Auflage
Bergisch Gladbach 1994, 208 S., 17,80 DM, ISBN 3-89012-366-X

Band 2: Eckart Bomsdorf
Induktive Statistik – Eine Einführung
6., durchgesehene Auflage
Bergisch Gladbach 1995, 216 S., 21,80 DM, ISBN 3-89012-426-7

REIHE: PLANUNG, ORGANISATION UND UNTERNEHMUNGSFÜHRUNG
Herausgegeben von Prof. Dr. Dr. h. c. Norbert Szyperski, Köln; Prof. Dr. Winfried Matthes, Wuppertal; Prof. Dr. Joachim Griese, Bern und Prof. Dr. Udo Winand, Kassel

Band 52: Jens Wallmann
Konversion als Unternehmensstrategie – Zentrale Problemfelder und Lösungsansätze der Umgestaltung von Hochtechnologie-Unternehmen mit verteidigungstechnischer Ausrichtung
Bergisch Gladbach 1994, 328 S., 67,— DM, ISBN 3-89012-393-7

Band 53: Markus Osburg
Einkaufsorganisation – Kriterien zur Organisation des Einkaufs in Konzernen der verarbeitenden Industrie
Bergisch Gladbach 1994, 308 S., 71,— DM, ISBN 3-89012-418-6

EINZELSCHRIFTEN

Susanne Rohloff
Die Unternehmungskultur im Rahmen von Unternehmungszusammenschlüssen
Bergisch Gladbach 1994, 268 S., 66,— DM, ISBN 3-89012-406-2

Matthias Eickhoff
Möglichkeiten und Grenzen bilanzanalytischer Erfolgsprognosen von Kapitalgesellschaften durch externe Jahresabschlußadressaten
Bergisch Gladbach 1994, 500 S., 89,— DM, ISBN 3-89012-411-9

Kai D. Kysela
Großhandelsmarketing
Bergisch Gladbach 1994, 412 S., 79,— DM, ISBN 3-89012-409-7

Andreas Otto
Unternehmenssteuerung im internationalen Wettbewerb – Theoretische Grundlagen und praktische Implikationen
Bergisch Gladbach 1994, 288 S., 68,— DM, ISBN 3-89012-403-8

Roland Dittmann
Entwicklung eines Expertensystems zur Beurteilung von Radiowerbung
Bergisch Gladbach 1994, 348 S., 74,— DM, ISBN 3-89012-416-X

Eckart Bomsdorf
Die ältere Bevölkerung des Freistaates Sachsen bis zum Jahr 2050 – Modellrechnungen unter besonderer Berücksichtigung eines landesspezifischen Rückgangs der Mortalität
Bergisch Gladbach 1995, 88 S., 29,— DM, ISBN 3-89012-425-9

Michael Ziegler
Deregulierung der Sonderabfallwirtschaft
Bergisch Gladbach 1995, 412 S., 79,— DM, ISBN 3-89012-431-3

Bruno Stegmüller
Internationale Marktsegmentierung als Grundlage für internationale Marketing-Konzeptionen
Bergisch Gladbach 1995, 392 S., 79,— DM, ISBN 3-89012-433-X

Frank Weingarten
Entlastung des Luftverkehrs in Deutschland unter den Bedingungen eines wachsenden Luftverkehrsmarktes
Bergisch Gladbach 1995, 324 S., 76,— DM, ISBN 3-89012-438-0

Doris Christophersen
Umfeldanalyse von Kunsthandels- und Kunstauktionsunternehmen – Ein Beitrag zur Ökonomie des Kunstmarktes
Bergisch Gladbach 1995, 324 S., 77,— DM, ISBN 3-89012-439-9

Theo Peters
Make-or-Buy-Entscheidungen im Marketingbereich
Bergisch Gladbach 1995, 264 S., 69,— DM, ISBN 3-89012-443-7

Peter Gunzenhauser
Unternehmenssanierung in den neuen Bundesländern – Eine Branchenuntersuchung des Werkzeugmaschinenbaus
Bergisch Gladbach 1995, 296 S., 74,— DM, ISBN 3-89012-449-6

REIHE: BIFOA-MONOGRAPHIEN
Herausgegeben von Prof. Dr. Dr. h. c. mult. Erwin Grochla †, Köln; Prof. Dr. Erich Frese, Köln und Prof. Dr. Dietrich Seibt, Köln

Band 32: Heiko Lippold, Heinz-Martin Hett, Jörg Hilgenfeldt, Dieter Klagge, Walter Nett
Elektronische Dokumentenverwaltung in Klein- und Mittelbetrieben
Bergisch Gladbach 1993, 152 S., 49,— DM, ISBN 3-89012-335-X

REIHE: BETRIEBLICHE PRAXIS

Band 1: Johannes Georg Bischoff/Jürgen Tracht
Wie mache ich mich als Handelsvertreter selbständig? 4. Auflage
Bergisch Gladbach 1992, 196 S., 58,— DM, ISBN 3-89012-306-6

Band 5: Johannes Georg Bischoff
Das Rechnungswesen der Handelsvertretung als Führungsinstrument
2. Auflage
Bergisch Gladbach 1991, 80 S., 29,— DM, ISBN 3-89012-254-X

Band 8: Johannes Georg Bischoff/Josef Breitbach/Ulrich Zelter (Hrsg.)
Der Schritt in die Selbständigkeit – Praktische Hinweise für Existenzgründer
5. Auflage
Bergisch Gladbach 1995, 232 S., 49,— DM, ISBN 3-89012-428-3

Band 9: Johannes Georg Bischoff (Hrsg.)
Der Schritt in die Selbständigkeit – Praktische Hinweise für Existenzgründer in Sachsen
2. Auflage
Bergisch Gladbach 1995, 252 S., 49,— DM, ISBN 3-89012-429-1

REIHE: KUNSTGESCHICHTE
Herausgegeben von Prof. Dr. Norbert Werner, Gießen

Band 7: Susanne Ließegang
HENRI MATISSE – Gegenstand und Bildrealität – Dargestellt an Beispielen der Malerei zwischen 1908 und 1918
Bergisch Gladbach 1994, 200 S., 53,— DM, ISBN 3-89012-382-1

REIHE: QUANTITATIVE ÖKONOMIE
Herausgegeben von Prof. Dr. Eckart Bomsdorf, Köln; Prof. Dr. Wim Kösters, Bochum und Prof. Dr. Winfried Matthes, Wuppertal

Band 55: Dominik Kramer
Kostenorientierte Reihenfolgeplanung
Bergisch Gladbach 1994, 360 S., 76,— DM, ISBN 3-89012-402-X

Band 56: Stefan Olbermann
Renditeunterschiede, Marktstruktur und dynamischer Wettbewerb
Bergisch Gladbach 1994, 244 S., 64,— DM, ISBN 3-89012-404-6

Band 57: Klaus Röder
Der DAX-Future – Bewertung und empirische Analyse
Bergisch Gladbach 1994, 196 S., 59,— DM, ISBN 3-89012-408-9

Band 58: Jürgen Groß
Schätzverfahren im allgemeinen gemischten linearen Modell
Bergisch Gladbach 1995, 188 S., 63,— DM, ISBN 3-89012-435-6

Band 59: Thomas König
Konstruktionsbegleitende Kalkulation auf der Basis von Ähnlichkeitsvergleichen
Bergisch Gladbach 1995, 276 S., 72,— DM, ISBN 3-89012-437-2

Band 60: Holger Claessen
Spezifikation und Schätzung von VARMA-Prozessen unter besonderer Berücksichtigung der Echelon-Form
Bergisch Gladbach 1995, 280 S., 72,— DM, ISBN 3-89012-447-X

Band 61: Peter Andreas Mrosik
Revision der Integration dynamischer industrieller Informationssysteme aus systemtheoretischer Sicht
Bergisch Gladbach 1995, 348 S., 79,— DM, ISBN 3-89012-448-8

Band 62: Andreas Südkamp
Einsatzmöglichkeiten quantitativer Prognoseverfahren im Rahmen der Betriebsergebnisplanung in der Eisen- und Stahlindustrie
Bergisch Gladbach 1995, 416 S., 86,— DM, ISBN 3-89012-451-8

Band 63: Josef Hünting
Rationale Erwartungen – Implikationen für Identifikation und Schätzung interdependenter ökonometrischer Modelle
Bergisch Gladbach 1995, 196 S., 64,— DM, ISBN 3-89012-454-2

Band 64: Udo Bankhofer
Unvollständige Daten- und Distanzmatrizen in der Multivariaten Datenanalyse
Bergisch Gladbach 1995, 240 S., 69,— DM, ISBN 3-89012-458-5

REIHE: VERSICHERUNGSWIRTSCHAFT
Herausgegeben von Prof. Dr. Dieter Farny, Köln

Band 16: Jürgen Riege
Gewinn- und Wachstumsstrategien von Versicherungsunternehmen
Bergisch Gladbach 1994, 416 S., 75,— DM, ISBN 3-89012-380-5

Band 17: Heiko Buck
Die versicherungstechnischen Rückstellungen im Jahresabschluß von Schaden- und Unfallversicherungsunternehmen – Nach Handels- und Ertragsteuerrecht unter besonderer Berücksichtigung der Versicherungstechnik
Bergisch Gladbach 1995, 364 S., 79,— DM, ISBN 3-89012-440-2

REIHE: WIRTSCHAFTSINFORMATIK
Herausgegeben von Prof. Dr. Dietrich Seibt, Köln; Prof. Dr. Dr. Ulrich Derigs, Köln und Prof. Dr. Werner Mellis, Köln

Band 12: Elke Schneider
Der Prozeß der Wissensakquisition und seine Integration in den Expertensystem-Entwicklungsprozeß
Bergisch Gladbach 1994, 444 S., 84,— DM, ISBN 3-89012-420-8

Band 13: Susanne Kirchhoff
Abbildungsqualität von wissensbasierten Systemen – Eine Methodologie zur Evaluierung
Bergisch Gladbach 1994, 432 S., 75,— DM, ISBN 3-89012-422-4

Band 14: Brigitte Reminger
Systematik der Erstellung, Validierung, organisatorischen Einbindung und Wirtschaftlichkeitsbetrachtung von Expertensystemen im Unternehmen
Bergisch Gladbach 1995, 296 S., 69,— DM, ISBN 3-89012-423-2

Band 15: Susanne Müller
Ablaufmodellierung als Analyse-, Entwurfs- und Realisierungsmethodik im Softwareentwicklungsprozeß
Bergisch Gladbach 1995, 336 S., 72,— DM, ISBN 3-89012-427-5

REIHE: PERSONAL-MANAGEMENT
Herausgegeben von Prof. Dr. Fred Becker, Jena und Prof. Dr. Jürgen Berthel, Siegen

Band 4: Frank Koslowski
Personalbezogene Frühaufklärung in Management und Controlling
Bergisch Gladbach 1994, 328 S., 72,— DM, ISBN 3-89012-415-1

Band 5: Andreas Ostmann
Personelle Implikationen des Management-Holding Konzepts – Strategische, organisatorische und rechtliche Einflußgrößen auf das Personalmanagement im Rahmen der Konzernführung
Bergisch Gladbach 1994, 320 S., 72,— DM, ISBN 3-89012-417-8

Band 6: Susanne Ribbert
Interim-Management durch externe Führungskräfte – Eine Analyse der Einsatzgebiete, Erfolgsdeterminanten und Gestaltungsmöglichkeiten
Bergisch Gladbach 1995, 272 S., 72,— DM, ISBN 3-89012-459-3

REIHE: EUROPÄISCHE WIRTSCHAFT
Herausgegeben von Prof. Dr. Winfried Matthes, Wuppertal

Band 3: Isabelle Hölper
Die Wettbewerbschancen der deutschen Süßwarenindustrie im EG-Binnenmarkt
Bergisch Gladbach 1994, 328 S., 72,— DM, ISBN 3-89012-410-0

Band 4: Manuela Schopp
Die Auswirkung der EU-Integration auf die Märkte für Butter und Butterersatzerzeugnisse in den neuen Bundesländern – Marktanalyse und Marketingmaßnahmen zur Förderung des Butterabsatzes
Bergisch Gladbach 1995, 328 S., 78,— DM, ISBN 3-89012-452-6

REIHE: INSTITUT FÜR BETRIEBLICHE DATENVERARBEITUNG (IBD) e.V., Fachhochschule Düsseldorf
Herausgegeben von Prof. Dr. Felicitas Albers

Band 1: Michael Hoppe
Organisation und DV-Unterstützung der Personalwirtschaft – Aufbauorganisatorische, ablauforganisatorische und informationstechnische Aspekte, dargestellt am Beispiel der Hauptverwaltung einer Forschungsgesellschaft
Bergisch Gladbach 1993, 112 S., 39,— DM, ISBN 3-89012-343-0

Band 2: Thomas Schleiken
Organisatorische Implementierungsprojekte – Am Beispiel der Einführung eines computergestützten Informationssystems in einem mittelständischen Unternehmen – Grundlagen, Schwachstellenanalyse und Gestaltungsempfehlungen
Bergisch Gladbach 1995, 128 S., 39,— DM, ISBN 3-89012-424-0

Band 3: Michael Bergmann
Textverarbeitung und Tabellenkalkulation mit dem PC
Ein Anwendungshandbuch zur Einführung in:
- Allgemeine Grundlagen
- Das Betriebssystem MS-DOS 6.2
- Die Benutzeroberfläche Windows 3.11
- Die Textverarbeitung Winword 6.0
- Die Tabellenkalkulation Excel 5.0

Bergisch Gladbach 1995, 260 S., 19,80 DM, ISBN 3-89012-453-4

REIHE: INTERNATIONALE WIRTSCHAFT
Herausgegeben von Prof. Dr. Manfred Borchert, Münster; Prof. Dr. Gustav Dieckheuer, Münster und Prof. Dr. Paul J. J. Welfens, Potsdam

Band 2: Monika Plum
Auswirkungen von Direktinvestitionen in Empfängerländern
Bergisch Gladbach 1995, 316 S., 76,— DM, ISBN 3-89012-434-8

Band 3: Markus Baumgartner
Internationale Auswirkungen einer nationalen Kapitalverknappung
Bergisch Gladbach 1995, 292 S., 74,— DM, ISBN 3-89012-436-4

Band 4: Jürgen Reckwerth
Theoretische und empirische Analyse internationaler Wirtschaftsbeziehungen mit einem Mehrländermodell
Bergisch Gladbach 1995, 260 S., 69,— DM, ISBN 3-89012-445-3

Band 5: Dirk Schallenberg
Akquisitionen und Kooperationen – Eine entscheidungsorientierte Analyse von Unternehmenszusammenschlüssen in der Textilwirtschaft
Bergisch Gladbach 1995, 300 S., 75,— DM, ISBN 3-89012-446-1

Band 6: Kirsten Witte
Ordnungspolitische Perspektiven der Europäischen Union – Eine Analyse aus Sicht der Neuen Institutionenökonomik
Bergisch Gladbach 1995, 212 S., 66,— DM, ISBN 3-89012-456-9

REIHE: MEDIZINISCHE FORSCHUNG
Herausgegeben von Prof. Dr. med. Victor Weidtman (em.), Köln

Band 7: Markus Reuber
Staats- und Privatanstalten in Irland – Irre, Ärzte und Idioten (1600–1900)
Bergisch Gladbach 1994, 232 S., 57,— DM, ISBN 3-89012-394-5

Band 8: Vinod Patel
Diabetic Retinopathy : Haemodynamic and Clinical Factors in the Pathogenesis
Bergisch Gladbach 1995, 288 S., 69,— DM, ISBN 3-89012-432-1

Band 9: Maria Bibiana Ristig
Urformen der Kinderheilkunst im Spiegel altdeutscher Waisenhäuser (1600–1800)
Bergisch Gladbach 1995, 248 S., 69,— DM, ISBN 3-89012-455-0

REIHE: SCHRIFTEN DES INSTITUTS FÜR WOHNUNGSRECHT UND WOHNUNGSWIRTSCHAFT AN DER UNIVERSITÄT ZU KÖLN
Herausgegeben von den Direktoren des Instituts

Band 54: Stephan Dirk Hinsche
Immobilienservice als Angebotserweiterung in der Immobilienwirtschaft
Bergisch Gladbach 1994, 124 S., 48,— DM, ISBN 3-89012-419-4

Die vorliegende Arbeit beschäftigt sich mit dem Problem unvollständiger Daten- und Distanzmatrizen in der Multivariaten Datenanalyse. Im Fall fehlender Werte können die herkömmlichen, auf vollständigen Daten- oder Distanzmatrizen basierenden Auswertungsmethoden nicht mehr unmittelbar zur Anwendung kommen. Somit ergibt sich die Notwendigkeit einer expliziten Berücksichtigung der fehlenden Werte im Rahmen der datenanalytischen Untersuchung. In dieser Arbeit werden dabei vor allem die Möglichkeiten einer Analyse des Mechanismus, der den fehlenden Daten zugrundeliegt, sowie die darauf aufbauenden Verfahren zur Behandlung unvollständiger Daten- und Distanzmatrizen vorgestellt.

In der Literatur existiert mittlerweile zwar eine Vielzahl von Einzelarbeiten, die sich mit dem Problem fehlender Daten beschäftigen, eine grundlegende Gesamtdarstellung ist jedoch nicht zu finden. Diese Arbeit verfolgt daher das Ziel, sowohl die aus der Literatur bekannten wie auch die daraus ableitbaren bzw. darüber hinaus denkbaren Lösungsansätze und Methoden umfassend und mathematisch orientiert darzustellen und in ein Gesamtkonzept zur Auswertung einer unvollständigen Daten- bzw. Distanzmatrix zu integrieren.

Die Arbeit richtet sich somit an Wissenschaftler und Praktiker, die mit dem Problem unvollständiger Daten- und Distanzmatrizen konfrontiert werden. Für diesen Personenkreis soll die grundlegende Vorgehensweise zur adäquaten Behandlung fehlender Daten ausführlich dargelegt und erörtert werden.

Udo Bankhofer, geboren am 12. März 1966 in Dachau; von 1987 bis 1992 Studium der Betriebswirtschaftslehre an der Universität Augsburg, Abschluß zum Diplom-Kaufmann im Mai 1992; seit Juni 1992 wissenschaftlicher Mitarbeiter am Lehrstuhl für Mathematische Methoden der Wirtschaftswissenschaften an der WISO-Fakultät der Universität Augsburg; Promotion zum Dr. rer. pol. im Juni 1995.

DM 69,—